Text Mining from User Generated Content

Ronen Feldman

Information Systems Department
School of Business Administration
Hebrew University, Jerusalem, ISRAEL
Ronen.feldman@huji.ac.il

Lyle Ungar

Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19103
ungar@cis.upenn.edu

Outline

- Intro to text mining
  - Information Retrieval (IR) vs. Information Extraction (IE)
- Information extraction (IE)
- Open IE/Relation Extraction
- Sentiment mining
- Wrap-up

Text Mining != Search

Information Retrieval

Find *Documents* matching the Query

- Actual information buried inside documents
- Long lists of documents

Information Extraction

Display *Information* relevant to the Query

- Extract Information from within the documents
- Aggregate over entire collection

Text Mining

Input

Documents

Output

Patterns

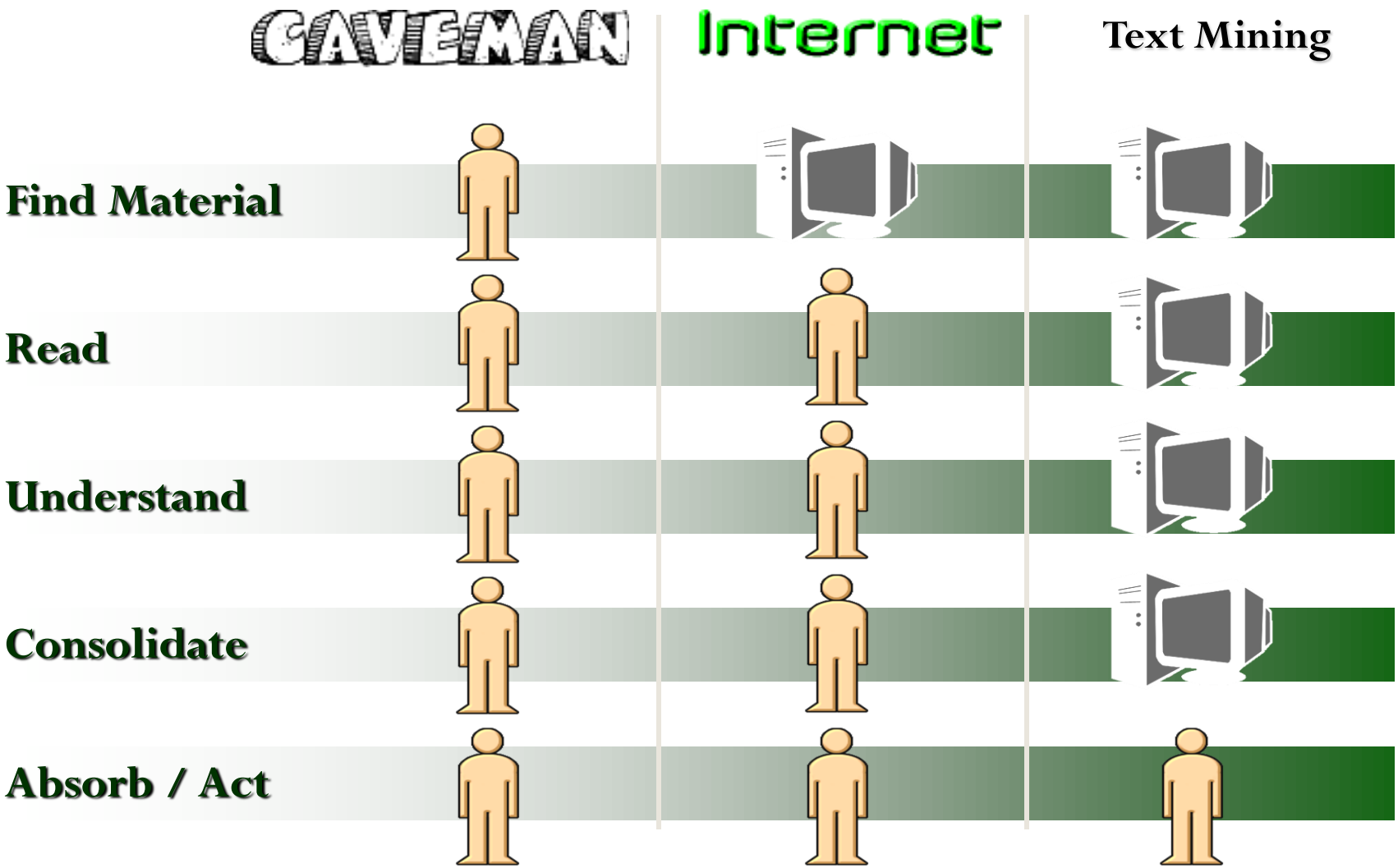
Connections

Profiles

Trends

Seeing the Forest for the Trees

Let text mining do the legwork for you



Outline

- Intro to text mining
- Information extraction (IE)
  - IE Components
- Open IE/Relation Extraction
- Sentiment mining
- Wrap-up

Information Extraction

Theory and Practice

What is Information Extraction?

- IE extracts pieces of information that are salient to the user's needs.
  - Find *named entities* such as persons and organizations
  - Find find *attributes* of those entities or *events* they participate in
  - Contrast IR, which indicates which documents need to be read by a user
- Links between the extracted information and the original documents are maintained to allow the user to reference context.

Applications of Information Extraction

- Infrastructure for IR and for Categorization
- Information Routing
- Event Based Summarization
- Automatic Creation of Databases
  - Company acquisitions
  - Sports scores
  - Terrorist activities
  - Job listings
  - Corporate titles and addresses

Why Information Extraction?

xerox

Type	Public (NYSE: XRX)
Founded	Rochester, New York, USA (1906)
Headquarters	Norwalk, Connecticut, USA Offices in Rochester, New York
Key people	Anne M. Mulcahy, Chairman & CEO Ursula Burns, President Larry Zimmerman, CFO Gary R. Kabureck CAO Michael MacDonald, President, Marketing Operations
Industry	Document Services Computer Peripherals
Products	Digital Imaging Printers
Revenue	▲$17.2 billion USD (2007)
Employees	57,400 (2007)
Website	www.xerox.com

"Who is the CEO of Xerox?"

"Female CEOs of public companies"

Text Sources

- Comments and notes
  - Physicians, Sales reps.
  - Customer response centers
  - Email
  - Word & PowerPoint documents
- Annotations in databases
  - e.g. GenBank, GO, EC, PDB
- The web
  - Blogs, newsgroups
- Newswire and journal articles
  - Medline has 13 million abstracts
- Facebook, tweets, search queries, …

Document Types

- Structured documents
  - Output from CGI
- Semi-structured documents
  - Seminar announcements
  - Job listings
  - Ads
- Free format documents
  - News
  - Scientific papers
  - Blogs, tweets, Facebook status, …

Relevant IE Definitions

- **Entity:** an object of interest such as a person or organization.
- **Attribute**: a property of an entity such as its name, alias, descriptor, or type.
- **Fact:** a relationship held between two or more entities such as the position of a person in a company.
- **Event:** an activity involving several entities such as a terrorist act, airline crash, management change, new product introduction.

IE Accuracy by Information Type

Information Type	Accuracy
Entities	90-98%
Attributes	80%
Facts	60-70%
Events	50-60%

IE input: free text

JERUSALEM - A Muslim suicide bomber blew apart 18 people on a Jerusalem bus and wounded 10 in a mirror-image of an attack one week ago. The carnage could rob Israel's Prime Minister Shimon Peres of the May 29 election victory he needs to pursue Middle East peacemaking. Peres declared all-out war on Hamas but his tough talk did little to impress stunned residents of Jerusalem who said the election would turn on the issue of personal security.

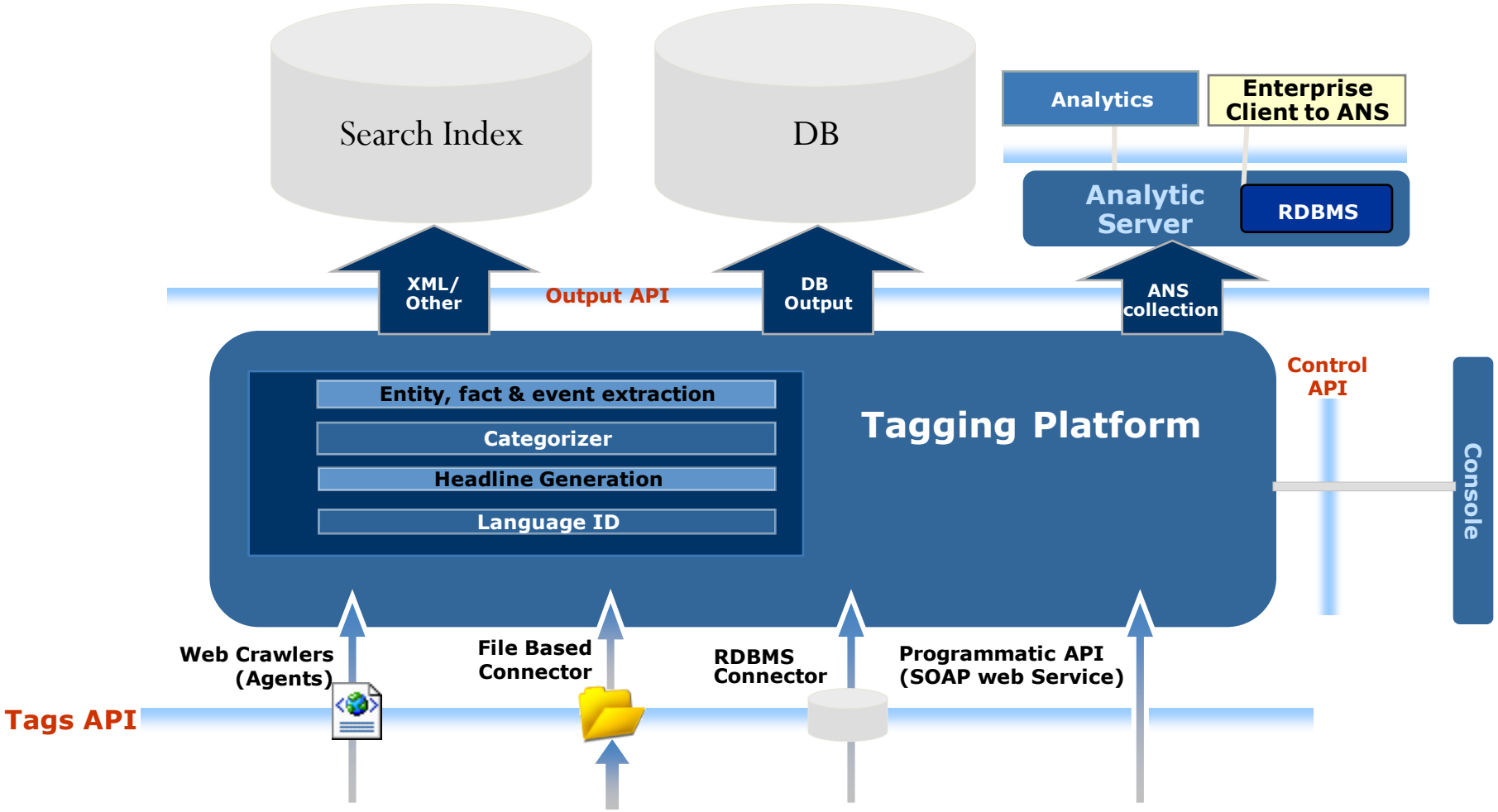
IE – Extracted Information

MESSAGE: ID	TST-REU-0001
SECSOURCE: SOURCE	Reuters
SECSOURCE: DATE	March 3, 1996, 11:30
INCIDENT: DATE	March 3, 1996
INCIDENT: LOCATION	Jerusalem
INCIDENT: TYPE	Bombing
HUM TGT: NUMBER	"killed: 18" "wounded: 10”
PERP: ORGANIZATION	"Hamas"

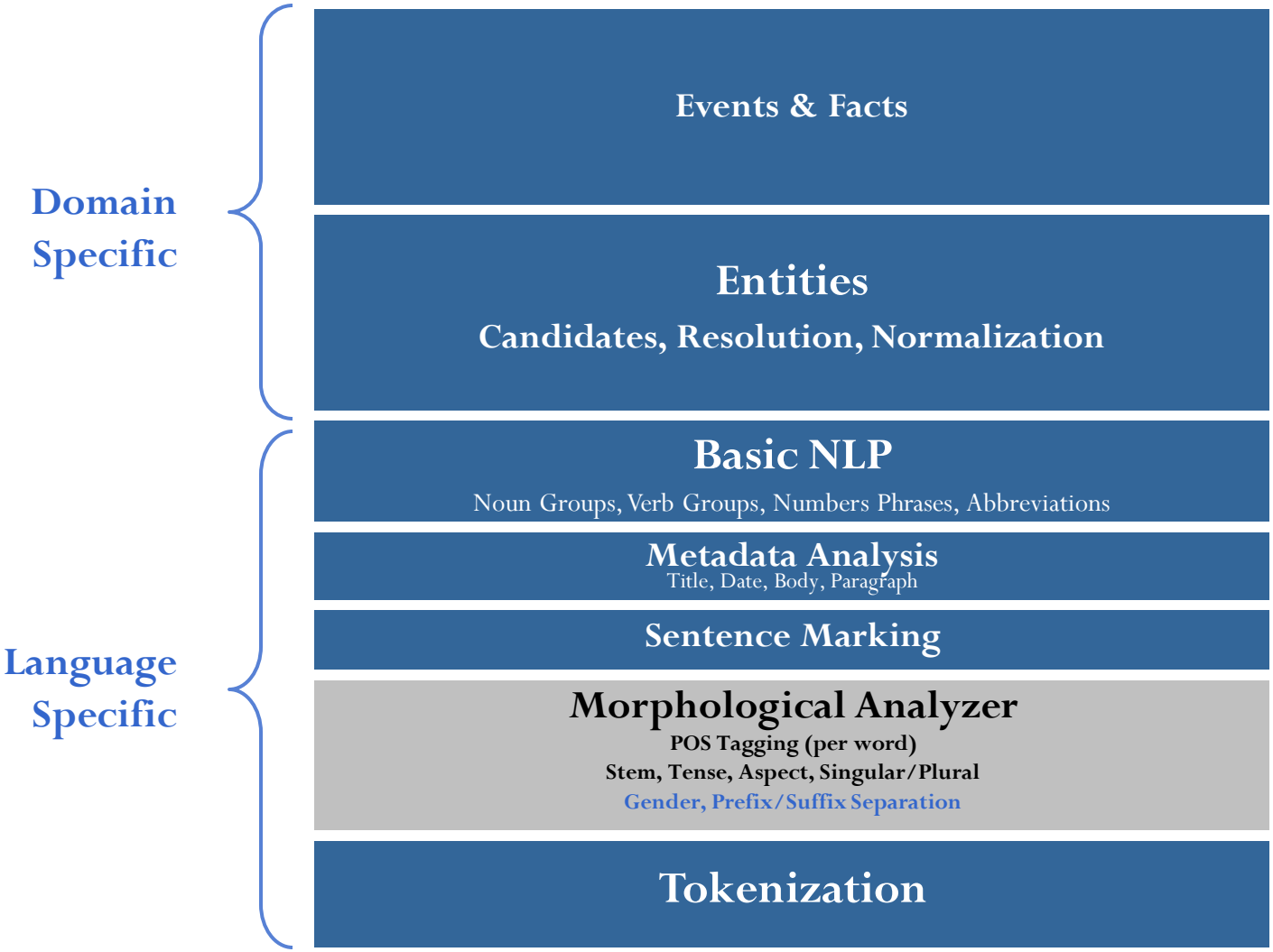
IE - Method

- Extract raw text (html, pdf, ps, gif)
- Tokenize
- Detect term boundaries
  - We extracted *alpha 1 type XIII collagen* from …
  - Their house council recommended …
- Detect sentence boundaries
- Tag parts of speech (POS)
  - *John*/noun *saw*/verb *Mary*/noun.
- Tag named entities
  - Person, place, organization, gene, chemical
- Parse
- Determine co-reference
- Extract knowledge

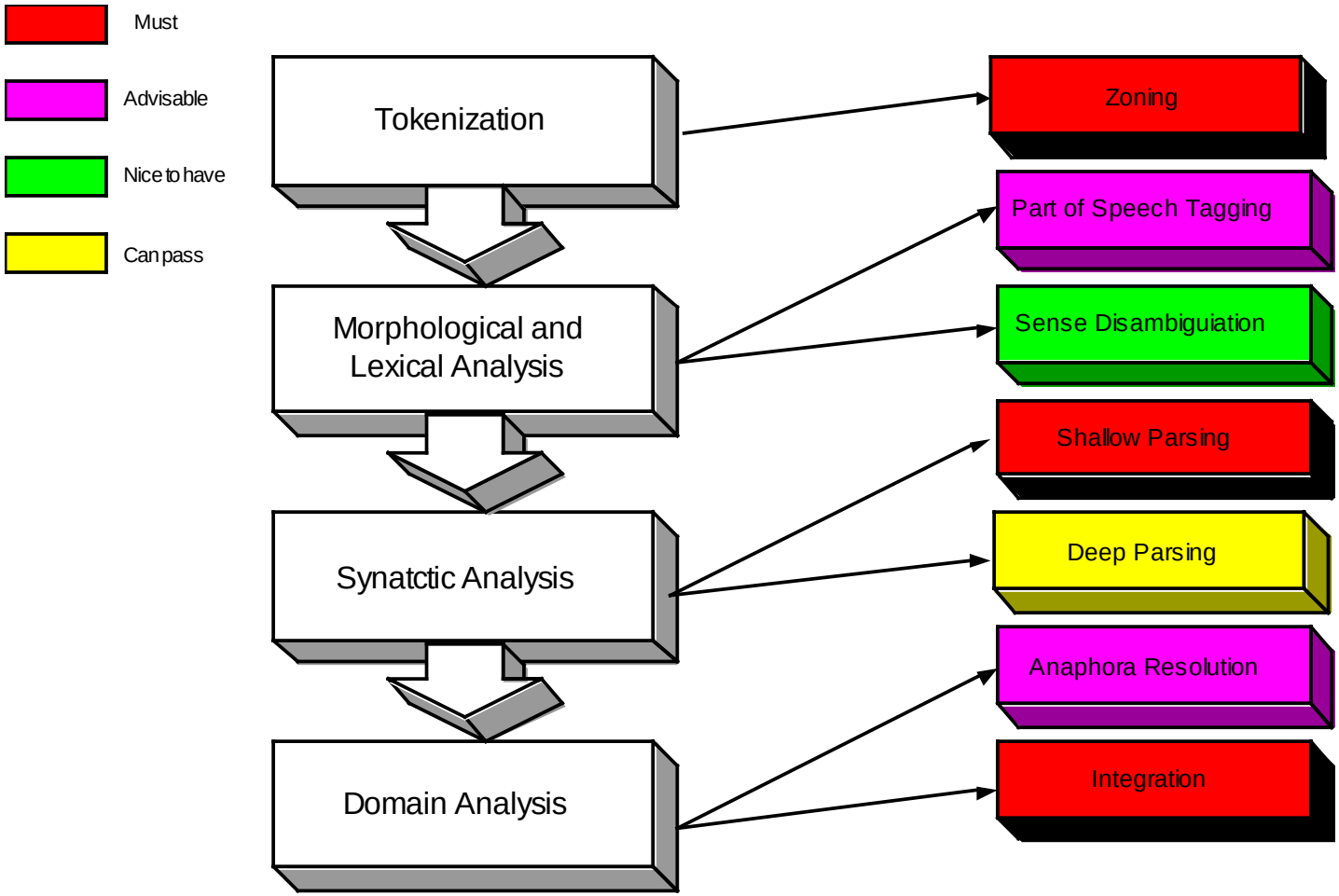
General Architecture



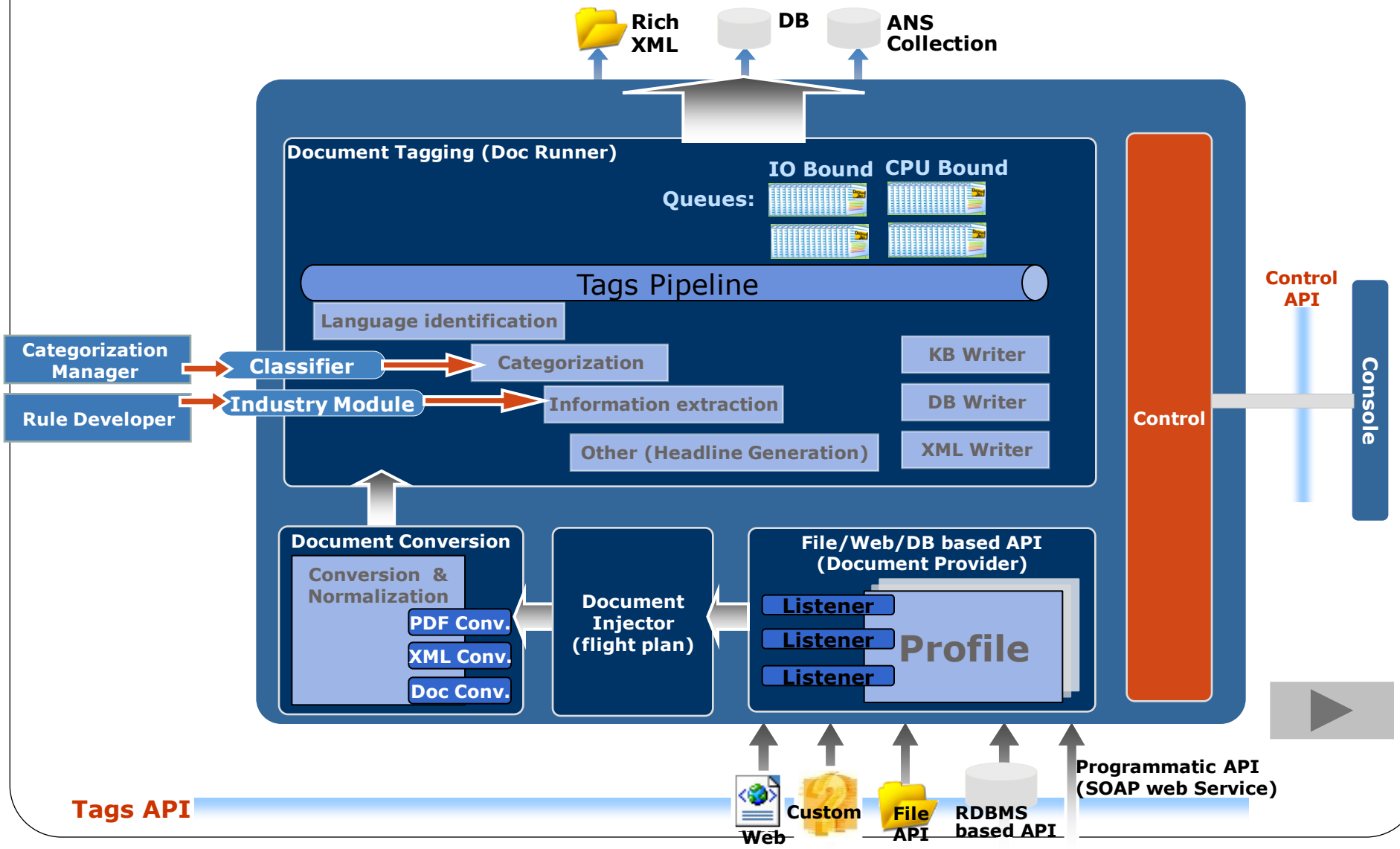
The Language Analysis Stack



Components of IE System



Tagging Architecture



Intelligent Auto-Tagging

(c) 2001, Chicago Tribune.
Visit the Chicago Tribune on the Internet at
http://www.chicago.tribune.com/
Distributed by Knight Ridder/Tribune
Information Services.
By Stephen J. Hedges and Cam Simpson

…….

The Finsbury Park Mosque is the center of radical Muslim activism in England. Through its doors have passed at least three of the men now held on suspicion of terrorist activity in France, England and Belgium, as well as one Algerian man in prison in the United States.

``The mosque's chief cleric, Abu Hamza al-Masri lost two hands fighting the Soviet Union in Afghanistan and he advocates the elimination of Western influence from Muslim countries. He was arrested in London in 1999 for his alleged involvement in a Yemen bomb plot, but was set free after Yemen failed to produce enough evidence to have him extradited. ."

……

<Facility>**Finsbury Park Mosque**</Facility>

<Country>**England**</Country>

<Country>**France** </Country>

<Country>**England**</Country>

<Country>**Belgium**</Country>

<Country>**United States**</Country>

<Person>**Abu Hamza al-Masri**</Person>

```
<PersonPositionOrganization>
  <OFFLEN OFFSET="3576"  LENGTH="33" />
  <Person>Abu Hamza al-Masri</Person>
  <Position>chief cleric</Position>
  <Organization>Finsbury Park Mosque</Organization>
</PersonPositionOrganization>
```

<City>**London**</City>

```
<PersonArrest>
  <OFFLEN OFFSET="3814" LENGTH="61" />
  <Person>Abu Hamza al-Masri</Person>
  <Location>London</Location>
  <Date>1999</Date>
  <Reason>his alleged involvement in a Yemen bomb
          plot</Reason>
</PersonArrest>
```

Business Tagging Example

SAP Acquires Virsa for Compliance Capabilities

By Renee Boucher Ferguson

April 3, 2006

Honing its software compliance skills, SAP announced April 3 the acquisition of Virsa Systems, a privately held company that develops risk management software.

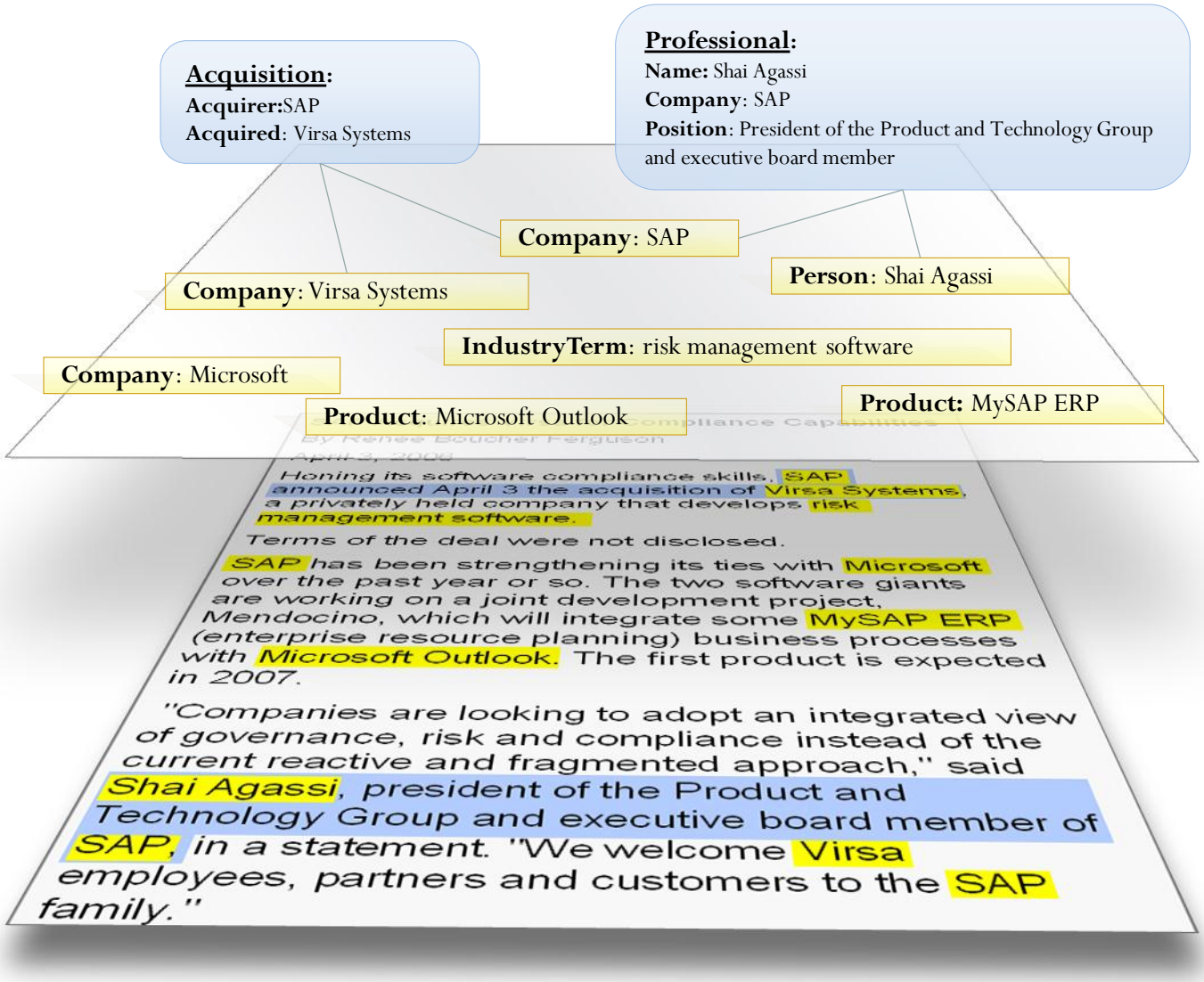
Terms of the deal were not disclosed.

SAP has been strengthening its ties with Microsoft over the past year or so. The two software giants are working on a joint development project, Mendocino, which will integrate some MySAP ERP (enterprise resource planning) business processes with Microsoft Outlook. The first product is expected in 2007.

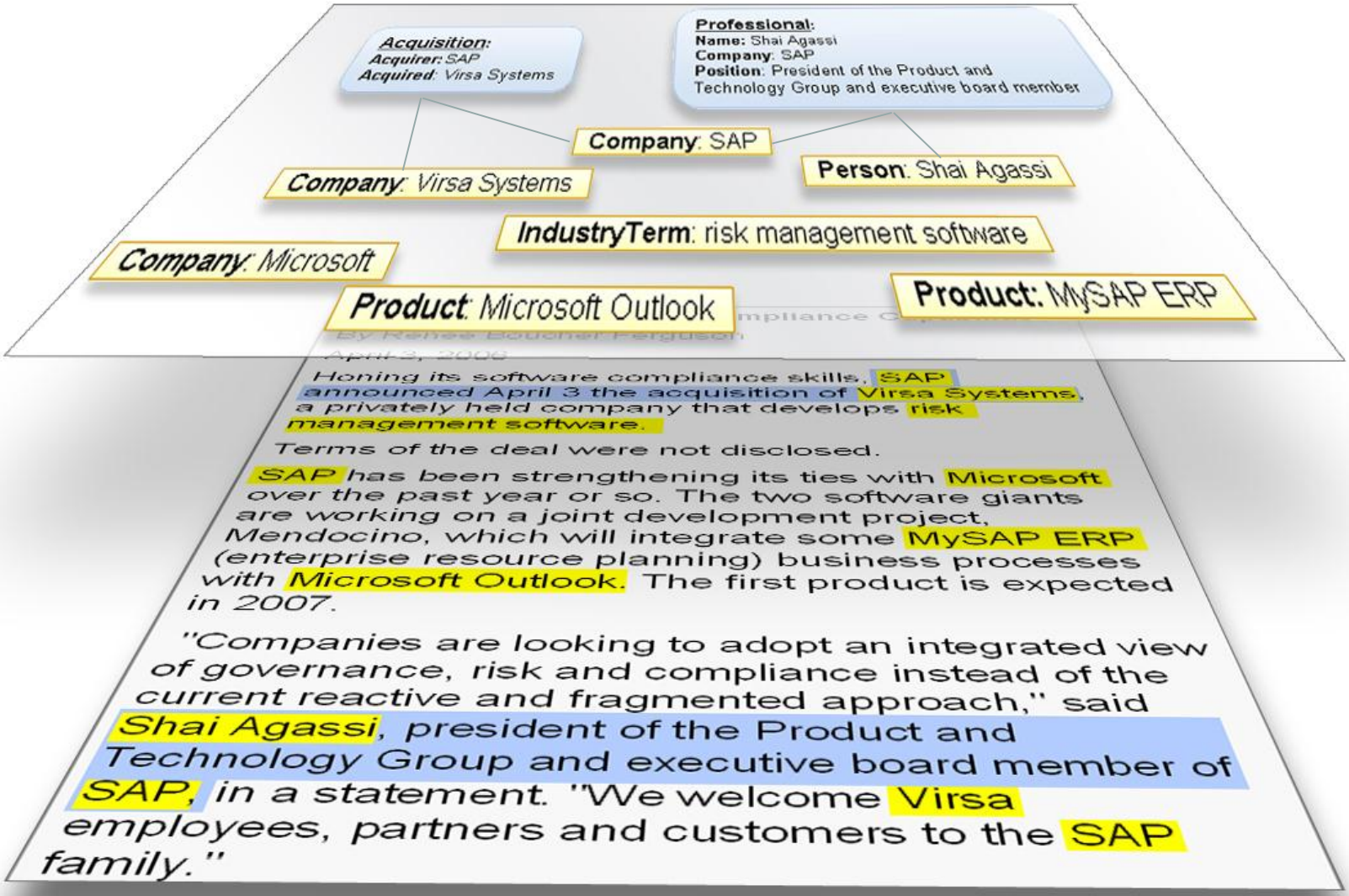
"Companies are looking to adopt an integrated view of governance, risk and compliance instead of the current reactive and fragmented approach," said Shai Agassi, president of the Product and Technology Group and executive board member of SAP, in a statement. "We welcome Virsa employees, partners and customers to the SAP family."

```
<Topic>BusinessNews</Topic>
<Company>SAP</Company>
<Company>Virsa Systems</Company>
<IndustryTerm>risk management software</IndustryTerm>
<Acquisition offset="494" length="130">
  <Company_Acquirer>SAP</Company_Acquirer>
  <Company_Acquired>Virsa Systems </Company_Acquired>
  <Status>known</Status>
</Acquisition>
<Company>SAP</Company>
<Company>Microsoft</Company>
<Product>MySAP ERP</Product>
<Product>Microsoft Outlook</Product>
<Person>Shai Agassi</Person>
<Company>SAP</Company>
<PersonProfessional offset="2789" length="92">
  <Person>Shai Agassi</Person>
  <Position>president of the Product and Technology Group
  and executive board member</Position>
  <Company>SAP</Company>
</PersonProfessional>
```

Business Tagging Example



Business Tagging Example



Leveraging Content Investment

Any type of content

- Unstructured textual content (current focus)
- Structured data; audio; video (future)

In any format

- Documents; PDFs; E-mails; articles; etc
- "Raw" or categorized
- Formal; informal; combination

From any source

- WWW; file systems; news feeds; etc.
- Single source or combined sources

Approaches for Building IE Systems: **Knowledge Engineering**

- Rules are crafted by linguists in cooperation with domain experts.
- Most of the work is done by inspecting a set of relevant documents.
- Can take a lot of time to fine tune the rule set.
- Best results were achieved with KB based IE systems.
- Skilled/gifted developers are needed.
- A strong development environment is a MUST!

IE – Templates (hand built)

<*victim*> was murdered

<*victim*> was killed

bombed <*target*>

bomb against <*target*>

killed with <*instrument*>

was aimed at <*target*>

offices in <*loc*>

operates in <*loc*>

facilities in <*loc*>

owned by <*company*>

<*company*> has positions

offices of <*company*>

Approaches for Building IE Systems: **Statistical Methods**

- The techniques are based on statistics
  - E.g., Conditional Random Fields (CRFs)
  - use almost no linguistic knowledge
  - are language independent
- The main input is an annotated corpus
  - Need a relatively small effort when building the rules, however creating the annotated corpus is extremely laborious.
  - Huge number of training examples is needed in order to achieve reasonable accuracy.
- Hybrid approaches can utilize the user input in the development loop.

Statistical Models

- Naive Bayes model:
  - generate class label y_i
  - generate word w_i from $Pr(W=w_i \mid Y=y_i)$
- Logistic regression: conditional version of Naive Bayes
  - set parameters to maximize $\sum_i \lg \Pr(y_i \mid x_i)$
- HMM model:
  - generate states $y_1,\ldots,y_n$ from $Pr(Y=y_i \mid Y=y_{i-1})$
  - generate words $w_1,\ldots,w_n$ from $Pr(W=w_i \mid Y=y_i)$
- Conditional version of HMMs
  - *Conditional Random Fields* (CRFs)

Conditional: estimate $p(y|x)$;
don't estimate $p(x)$

What Is Unique in Text Mining?

- Feature extraction.
- Very large number of features that represent each of the documents.
- The need for background knowledge.
- Even patterns supported by small number of document may be significant.
- Huge number of patterns, hence need for visualization, interactive exploration.

Language is complex!

Text Representations (Features)

- Character Trigrams
- Words/Parts of Speech
- Terms, Entities
- Linguistic Phrases
- Parse trees
- Relations, Frames, Scripts

Text mining is hard: Language is complex

- Synonyms and Orthonyms
  - Bush, HEK
- Anaphora (and Sortal anaphoric noun phrases)
  - It, they, the protein, both enzymes
- Notes are rarely grammatical
- Complex structure
  - The first time I bought your product, I tried it on my dog, who became very unhappy and almost ate my cat, who my daughter dearly loves, and then when I tried it on her, she turned blue!

Text mining is hard

- **Hand-built systems give poor coverage**
  - Large vocabulary
    - Chemicals, genes, names
  - Zipf's law
    - activate is common;
    - colocalize and synergize
    - are not
    - Most words are very rare
  - Can't manually list all patterns
- **Statistical methods need training data**
  - Expensive to manually label data

Text mining is easy

- Lots of redundant data
- Some problems are easy
  - IR: bag of words works embarrassingly well
  - Latent Semantic Analysis (LSA/SVD) for grading tests
- Incomplete, inaccurate answers often useful
  - Exploratory Data Analysis (EDA)
    - Suggest trends or linkages

Conclusions

- What doesn't work
  - Anything requiring high precision, broad coverage, and full automation
- What does work
  - Text mining with humans "in the loop"
    - Information retrieval, Message routing
    - Trend spotting
  - Specialized extractors
    - Company addresses, Sports scores …
- What will work
  - Using extracted info in statistical models
  - Speech to text

The Bottom Line

- Information extraction works great if you can afford to be 90% accurate
  - Generally requires human post-processing for > 95%
    - Unless the system is very highly specialized

Outline

- Intro to text mining
- Information extraction (IE)
- Open IE/Relation Extraction
  - Basic Open IE: TextRunner
  - Advanced Open IE: KnowItAll and SRES
- Sentiment mining
- Wrap-up

Relation Extraction

"Oh dear."

IE for the Web

Challenges

Difficult, ungrammatical sentences

Unreliable information

Heterogeneous corpus

Massive Number of Relations

Advantages

"Semantically tractable" sentences

Redundancy

Open IE
[Banko, *et al.* 2007]

TextRunner Search

http://www.cs.washington.edu/research/textrunner/

TextRunner Search

[Banko et al., 2007]

Retrieved **2760** results for **What kills bacteria?**.

Grouping results by predicate. Group by: argument 1 | argument 2

kills - 42 results

strong antibiotics (103), Antibiotics (67), Benzoyl peroxide (50), ***175 more...*** **kills** **bacteria**
Ultraviolet disinfection devices (3), ozone (3), iodine (2), ***7 more...*** may **kill** **bacteria** and viruses
Levaquin (21) **kills** a variety of **bacteria**
INH (4), the medicine (4) **kills** the TB **bacteria**
many antibiotics (3), Antibiotics (2), the " bad " bacteria (2) also **kills** the " good " **bacteria**
Infact Doxy (4), only the Doxy (2) **kills** a whole bunch of various **bacteria**
Treatment (4), Penicillin treatment (2) will **kill** the syphilis bacterium
SILVER (3), our disinfectant solution (2) **kills** almost all known **bacteria**
boiling (2), boil-water alerts (2) will **kill** **bacteria** and parasites
a food (2), antibiotics (2) can **kill** all **bacteria**
Anti-bacterial cleaners (4) **kills** 99.9 % of **bacteria** Cleans appliances
Appropriate treatment (4) **kills** the Shigella **bacteria**
artemisinin (3) can **kill** other parasites and **bacteria**
the chlorine dioxide (3) **kills** the already formed **bacteria**
this mouthwash (3) **kills** germs and **bacteria**
those drugs (3) **killed** Andrew 's normal gut-protective **bacteria**
Antibiotics (3) **kill** gonorrhea **bacteria**
Proper cooking (3) **kills** food poisoning **bacteria**
that microwaves (2) can **kill** the anthrax **bacteria**
Hot dry vapor steam (2) **kills** mold , mildew , viruses , **bacteria**
One application (2) **kills** **bacteria** odors
Benzoyl peroxide (2) **kills** off **bacteria**
Iodine (2) will **kill** the lactic **bacteria**
the boiling (2) **kills** impurities and **bacteria**
The chlorine (2) **kills** iron **bacteria**
ozone (2) **kills** the acid producing **bacteria**
Ampicillin (2) **kills** susceptible **bacteria**
Any positively offset frequency (2) **kills** all **bacteria** , viruses and parasites

Find: Previous Next Highlight all Match case

Transferring data from turingc.cs.washington.edu...

does not kill - 1 result

Doxycycline (14), Freezing (11), Refri

to kill - 6 results

antibiotics (7), water (3), milk (3), **2 n**

antibiotics (2) to **kill** extracellular **bac**

the ability (2) to **kill** a wide variety of

milk (2) to **kill** harmful **bacteria**

a second time (2) to **kill** any **bacteria**

macrophages (2) to **kill** the intracellu

helps kill - 2 results

Raw garlic (2), lime juice (2), uv germ

Benzoyl peroxide (3) helps **kill** skin **b**

does n't kill - 1 result

Freezing (6), Irradiation (4), antacids (2) does n't **kill** **bacteria**

kill not only - 1 result

Antibiotics (6), these drugs (3) **kill** not only harmful **bacteria**

x

Refrigeration and freezing do not kill bacteria, but slow th...

Refrigeration does not kill bacteria and can not improve food quality.

Refrigeration and freezing do not kill bacteria, but s...

Refrigeration and freezing do not kill bacteria, but only slow their growth.

Refrigeration and freezing do not kill bacteria, but sl...

Refrigeration does not kill most bacteria.

Refrigeration and freezing do not kill bacteria, but slow their growth.

Remember: refrigeration does not kill bacteria; it only slows down their growth .

TextRunner

[Banko, Cafarella, Soderland, *et al.*, IJCAI '07]

TextRunner Search Results

http://turingc.cs.washington.edu:7125/TextRunner/cgi-bin/wikirunner-hyp.pl

TextRunner Search

100-million page corpus

Retrieved **1113** results for **alan turing** in argument 1.

Grouping results by argument 1. Group by: predicate | argument 2

Alan Turing - 34 results

Alan Turing was British mathematician (8), founder of computer science (4), cryptographer (3), ***7 more...***
Alan Turing publishes paper (4), Intelligence (3), article (2), ***2 more...***
Alan Turing proposed test (4), Turing Test (4), Turing machine (2)
Alan Turing was born in London (4), Paddington (2), nursing home (2)
Alan Turing made appearance (3), foundational contributions (2)
Alan Turing committed suicide (4)
Alan Turing invented Turing Test (2), computers (2)
Alan Turing came up with idea (4)
Alan Turing is mathematician (2), many early pioneer (2)
Alan Turing introduced Turing Test (3)
Alan Turing was born June 23 , 1912 (3)
Alan Turing died in 1954 (3)
Alan Turing lived at Crown Inn (2)
Alan Turing was involved in code breaking activites (2)
Alan Turing lodged at Crown Inn (2)
Alan Turing contributed to Church_Turing_Deutsch principle (2)
Alan Turing broke Nazi code (2)

Search again:

Argument 1
alan turing
Predicate
Argument 2
Search

Jump to:

Alan Turing (34)
Alan Mathison **Turing** (3)
the **Alan Turing** Institute (2)
Alan Turing homepage (1)
The death of **Alan Turing** (1)
Alan M. **Turing** (1)
The **Alan Turing** Memorial (1)
the novel **Alan Turing** (1)
the British mathematician **Alan Turing** (1)
Dr. **Alan** M. **Turing** (1)
A young **Alan Turing** (1)
Alan Turing 's person element (1)

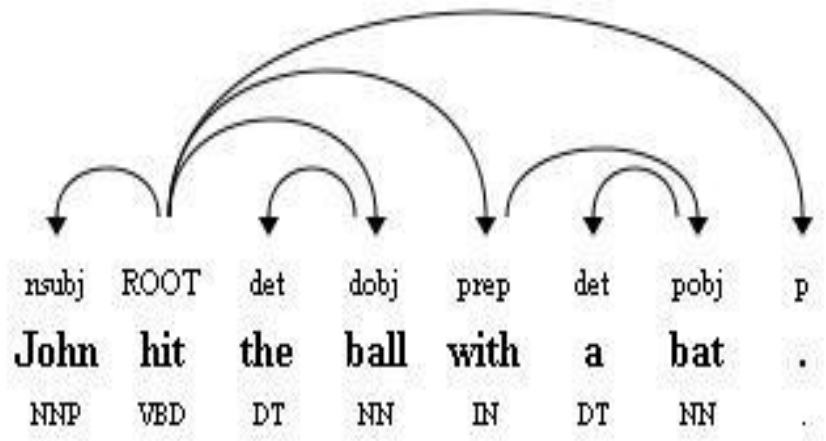
Done

Open IE

- Relation-Independent Extraction
  - How are relations expressed, in general?
  - Unlexicalized
- Self-Supervised Training
  - Automatically label training examples
- Discover relations on the fly
    - Traditional IE: $(e_1, e_2) \in R$?
    - Open IE: ***<u>What</u> is R?***

Training

- No parser at extraction time
- Use trusted parses to auto-label training examples
- Describe instances without parser-based features
  - Unlexicalized ⮝ PennTreeBank OK



+ (John, hit, ball)
+ (John, hit with, bat)
- (ball, with, bat)

Features

- Unlexicalized
  - Closed class words OK
- Parser-free
  - Part-of-speech tags, phrase chunk tags
  - ContainsPunct, StartsWithCapital, …
- Type-independent
  - Proper vs. common noun, no NE types

Relation Discovery

- Many ways to express one relation
- Resolver [Yates & Etzioni, HLT '07]

(Viacom, **acquired**, Dreamworks)
(Viacom, **'s acquisition of**, Dreamworks)
(Viacom, **sold off**, Dreamworks)

(Google, **acquired**, YouTube)
(Google Inc., **'s acquisition of**, YouTube)

(Adobe, **acquired**, Macromedia)
(Adobe, **'s acquisition of**, Macromedia)

$$P(R_1 = R_2) \sim \text{shared objects} * \text{strSim}(R_1, R_2)$$

Traditional IE vs. Open IE

	Traditional IE	**Open IE**
Input	Corpus + Relations + Training Data	Corpus + Relation-Independent Heuristics
Relations	Specified in Advance	Discovered Automatically
Features	Lexicalized, NE-Types	Unlexicalized, No NE types

Questions

- How does OIE fare when the relation set is unknown?
- Is it even possible to learn relation-independent extraction patterns?
- How do OIE and Traditional IE compare when the relation is given?

Eval 1: Open Info. Extraction (OIE)

- CRF gives better recall than Naïve Bayes (NB) Classifiers
  - Apply to 500 sentences from Web IE training corpus [Bunescu & Mooney '07]
  - P = precision, R = Recall, F1 = 2 P R/(P+R)

OIE-NB			OIE-CRF		
P	R	F1	P	R	F1
86.6	23.2	36.6	**88.3**	**45.2**	**59.8**

Category	Pattern	RF
Verb	E_1 Verb E_2 *X established Y*	37.8
Noun+Prep	E_1 NP Prep E_2 *the X* settlement with *Y*	22.8
Verb+Prep	E_1 Verb Prep E_2 X moved to Y	16.0
Infinitive	E_1 to Verb E_2 *X to acquire Y*	9.4
Modifier	E_1 Verb E_2 NP *X is Y winner*	5.2
Coordinate$_n$	E_1 (and|,|-|:) E_2 NP *X - Y deal*	1.8
Coordinate$_v$	E_1 (and|,) E_2 Verb *X , Y merge*	1.0
Appositive	E_1 NP (:|,)? E_2 *X hometown : Y*	0.8

Relation-Independent Patterns

- 95% could be grouped into 1 of 8 categories
- Dangerously simple
  - `Paramount , the` **`Viacom`** `- owned studio , bought` **`Dreamworks`**
  - **`Charlie Chaplin`** `, who died in 1977 , was born in` **`London`**
- Precise conditions
  - Difficult to specify by hand
  - Learnable by OIE model

Results

Category	OIE-NB			OIE-CRF		
	P	R	F1	P	R	F1
Verb	100.0	38.6	55.7	93.9	65.1	76.9
Noun+Prep	100.0	9.7	17.5	89.1	36.0	51.2
Verb+Prep	95.2	25.3	40.0	95.2	50.0	65.6
Infinitive	100.0	25.5	40.7	95.7	46.8	62.9
Other	0	0	0	0	0	0
All	86.6	23.2	36.6	**88.3**	**45.2**	**59.8**

Open IE is good at identifying verb and some noun-based relationships; others are hard because they are based on punctuation

Traditional IE with R1-CRF

- Trained from hand-labeled data per relation
- Lexicalized features, same graph structure
- Many relation extraction systems do this

  [e.g. Bunescu ACL '07, Culotta HLT '06]

- Question: what is effect of
  - Relation-specific/independent features
  - Supervised vs. Self-supervised Training

  keeping underlying models equivalent

Eval 2: Targeted Extraction

- Web IE corpus from [Bunescu 2007]
  - Corporate-acquisitions (3042)
  - Birthplace (1853)
- Collected two more relations in same manner
  - Invented-Product (682)
  - Won-Award (354)
- Labeled examples by hand

Results

	R1-CRF			OIE-CRF	
Relation	**P**	**R**	**Train Ex**	**P**	**R**
Acquisition	67.6	69.2	3,042	75.6	19.5
Birthplace	92.3	64.4	1,853	90.6	31.1
InventorOf	81.3	50.8	682	88.0	17.5
WonAward	73.6	52.8	354	62.5	15.3
All	73.9	**58.4**	5,931	**75.0**	18.4

Open IE can match precision of supervised IE ***without***

- Relation-specific training
- 100s or 1,000s of examples *per relation*

Summary

- Open IE
  - High-precision extractions without cost of per-relation training
  - Essential when number of relations is large or unknown
- May prefer Traditional IE when
  - High recall is necessary
  - For a small set of relations
  - *And* can acquire labeled data
- Try it!

  http://www.cs.washington.edu/research/textrunner

Outline

- Intro to text mining
- Information extraction (IE)
- Open IE/Relation Extraction
  - Basic Open IE: TextRunner
  - Advanced methods: KnowItAll and SRES
- Sentiment mining
- Wrap-up

Self-Supervised Relation Learning from the Web

KnowItAll (KIA)

- Developed at University of Washington by Oren Etzioni and colleagues
  - (Etzioni, Cafarella et al. 2005).
- Autonomous, domain-independent system that extracts facts from the Web.
  - The primary focus of the system is on extracting entities (unary predicates), although KnowItAll is able to extract relations (N-ary predicates) as well.
- **Input** is a set of entity classes to be extracted, such as "city", "scientist", "movie", etc.,
- **Output** is a list of entities extracted from the Web.

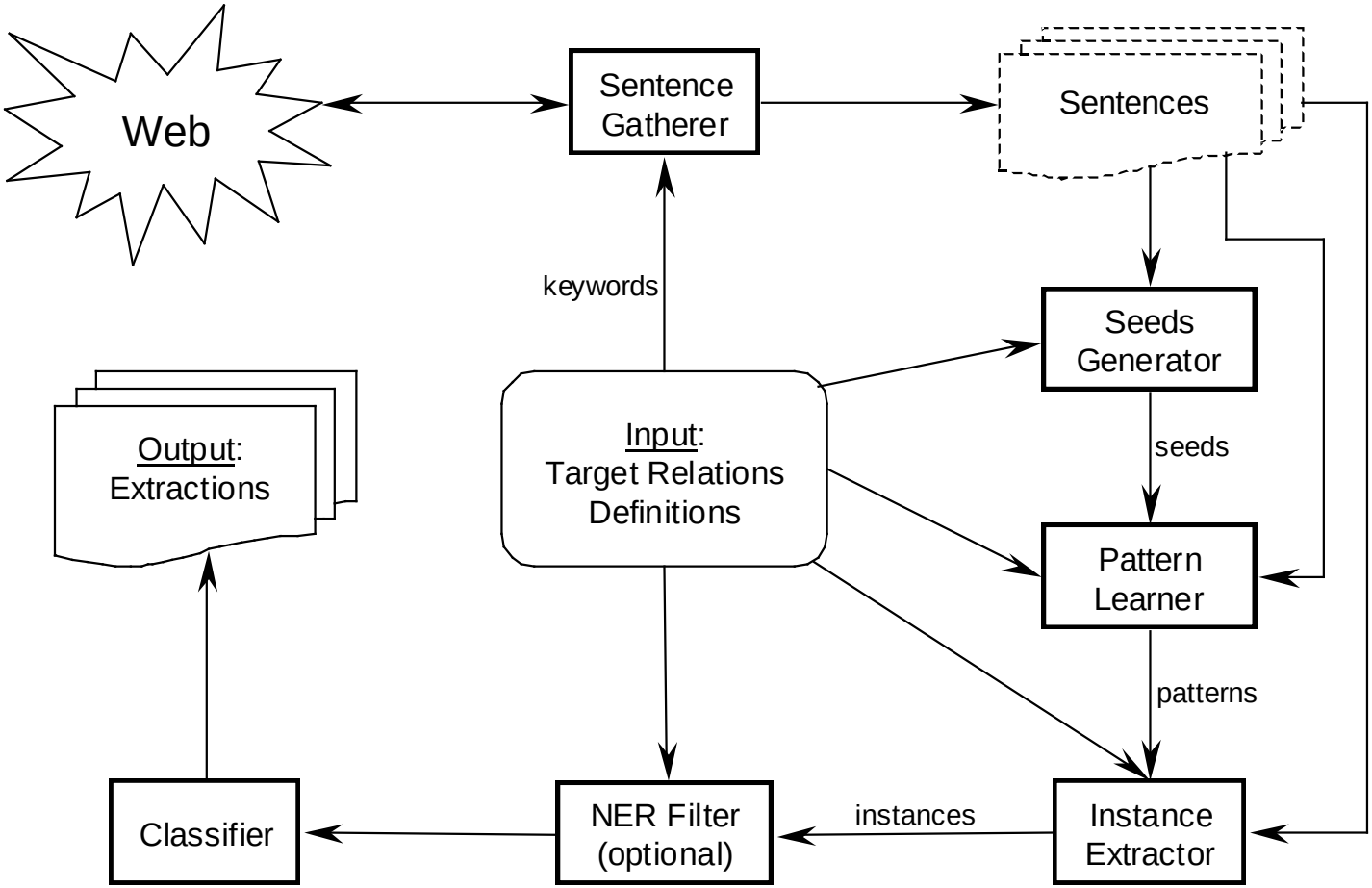
KnowItAll's Relation Learning

- The base version uses hand written patterns based on a general Noun Phrase (NP) tagger.
- The patterns used for extracting instances of
  - the ***Acquisition*(*Company*, *Company*)** relation:
    - NP2 "was acquired by" NP1
    - NP1 "'s acquisition of" NP2
  - the ***MayorOf*(*City*, *Person*)** relation:
    - NP ", mayor of" <city>
    - <city> "'s mayor" NP
    - <city> "mayor" NP

SRES

- SRES (**Self-Supervised Relation Extraction System**
  - learns to extract relations from the web in an unsupervised way.
  - takes as input the name of the relation and the types of its arguments
    - And a set of "seed" examples
  - Generates positive and negative examples
  - returns as output a set of extracted instances of the relation

SRES Architecture



Seeds for Acquisition

- Oracle – PeopleSoft
- Oracle – Siebel Systems
- PeopleSoft – J.D. Edwards
- Novell – SuSE
- Sun – StorageTek
- Microsoft – Groove Networks
- AOL – Netscape
- Microsoft – Vicinity
- San Francisco-based Vector Capital – Corel
- HP – Compaq

Positive Instances

- The positive set of a predicate consists of sentences that contain an instance of the predicate, with the actual instance's attributes changed to "*<AttrN>*", where *N* is the attribute index.
- For example, the sentence
  - *"The Antitrust Division of the U.S. Department of Justice evaluated the likely competitive effects of Oracle's proposed acquisition of PeopleSoft."*
- will be changed to
  - *"The Antitrust Division… …….effects of <Attr1>'s proposed acquisition of <Attr2>."*

Negative Instances

- Change the assignment of one or both attributes to other suitable entities in the sentence.
- In the shallow parser based mode of operation, any suitable noun phrase can be assigned to an attribute.

Examples

- ***The Positive Instance***
  - *"The Antitrust Division of the U.S. Department of Justice evaluated the likely competitive effects of <Attr1>'s proposed acquisition of <Attr2>"*
- ***Possible Negative Instances***
  - *<Attr1> of the <Attr2> evaluated the likely…*
  - *<Attr2> of the U.S. … …acquisition of <Attr1>*
  - *<Attr1> of the U.S. … …acquisition of <Attr2>*
  - *The Antitrust Division of the <Attr1> ….. acquisition of <Attr2>"*

Pattern Generation

- The patterns for a predicate *P* are generalizations of pairs of sentences from the positive set of *P*.
- The function *Generalize*(*S*1, *S*2) is applied to each pair of sentences *S*1 and *S*2 from the positive set of the predicate. The function generates a pattern that is the best (according to the objective function defined below) generalization of its two arguments.
- The following pseudo code shows the process of generating the patterns:

```
For each predicate P
    For each pair S1, S2 from PositiveSet(P)
        Let Pattern = Generalize(S1, S2).
        Add Pattern to PatternsSet(P).
```

Example Pattern Alignment

- *S1* = "*Toward this end, <Arg1> in July acquired <Arg2>*"
- *S2* = "*Earlier this year, <Arg1> acquired <Arg2>*"
- After the dynamical programming-based search, the following match will be found:

Toward		(cost 2)
	Earlier	(cost 2)
this	*this*	(cost 0)
end		(cost 2)
	year	(cost 2)
,	,	(cost 0)
<Arg1 >	*<Arg1 >*	(cost 0)
in July		(cost 4)
acquired	*acquired*	(cost 0)
<Arg2 >	*<Arg2 >*	(cost 0)

Generating the Pattern

- at total cost = 12. The match will be converted to the pattern
  - * * *this* * * , *<Arg1>* * *acquired* *<Arg2>*
- which will be normalized (after removing leading and trailing skips, and combining adjacent pairs of skips) into
  - *this* * , *<Arg1>* * *acquired* *<Arg2>*

Post-processing, filtering, and scoring of patterns

- **Remove** from each pattern all function words and punctuation marks that are surrounded by skips on both sides.
  - Thus, the pattern
    - *this * , <Arg1> * acquired <Arg2>*
  - from the example above will be converted to

    *, <Arg1> * acquired <Arg2>*

Content Based Filtering

- Every pattern must contain at least one word relevant (defined via WordNet) to its predicate.
- For example, the pattern

  *<Arg1> * by <Arg2>*
- will be removed, while the pattern

  *<Arg1> * purchased <Arg2>*
- will be kept, because the word "*purchased*" can be reached from "*acquisition*" via synonym and derivation links.

Scoring the Patterns

- Score the filtered patterns by their performance on the positive and negative sets.

Sample Patterns - Inventor

- X , .* inventor .* of Y
- X invented Y
- X , .* invented Y
- when X .* invented Y
- X ' s .* invention .* of Y
- inventor .* Y , X
- Y inventor X
- invention .* of Y .* by X
- after X .* invented Y
- X is .* inventor .* of Y
- inventor .* X , .* of Y
- inventor of Y , .* X ,
- X is .* invention of Y
- Y , .* invented .* by X
- Y was invented by X

Sample Patterns – CEO (Company/X,Person/Y)

- X ceo Y
- X ceo .* Y ,
- former X .* ceo Y
- X ceo .* Y .
- Y , .* ceo of .* X ,
- X chairman .* ceo Y
- Y , X .* ceo
- X ceo .* Y said
- X ' .* ceo Y
- Y , .* chief executive officer .* of X
- said X .* ceo Y
- Y , .* X ' .* ceo
- Y , .* ceo .* X corporation
- Y , .* X ceo
- X ' s .* ceo .* Y ,
- X chief executive officer Y
- Y , ceo .* X ,
- Y is .* chief executive officer .* of X

Score Extractions using a Classifier

- Score each extraction using the information on the instance, the extracting patterns and the matches.
- Assume extraction *E* was generated by pattern *P* from a match *M* of the pattern *P* at a sentence *S*. The following properties are used for scoring:
  1. Number of different sentences that produce *E* (with any pattern).
  2. Statistics on the pattern *P* generated during pattern learning – the number of positive sentences matched and the number of negative sentences matched.
  3. Information on whether the slots in the pattern *P* are anchored.
  4. The number of non-stop words the pattern *P* contains.
  5. Information on whether the sentence *S* contains proper noun phrases between the slots of the match *M* and outside the match *M*.
  6. The number of words between the slots of the match *M* that were matched to skips of the pattern *P*.

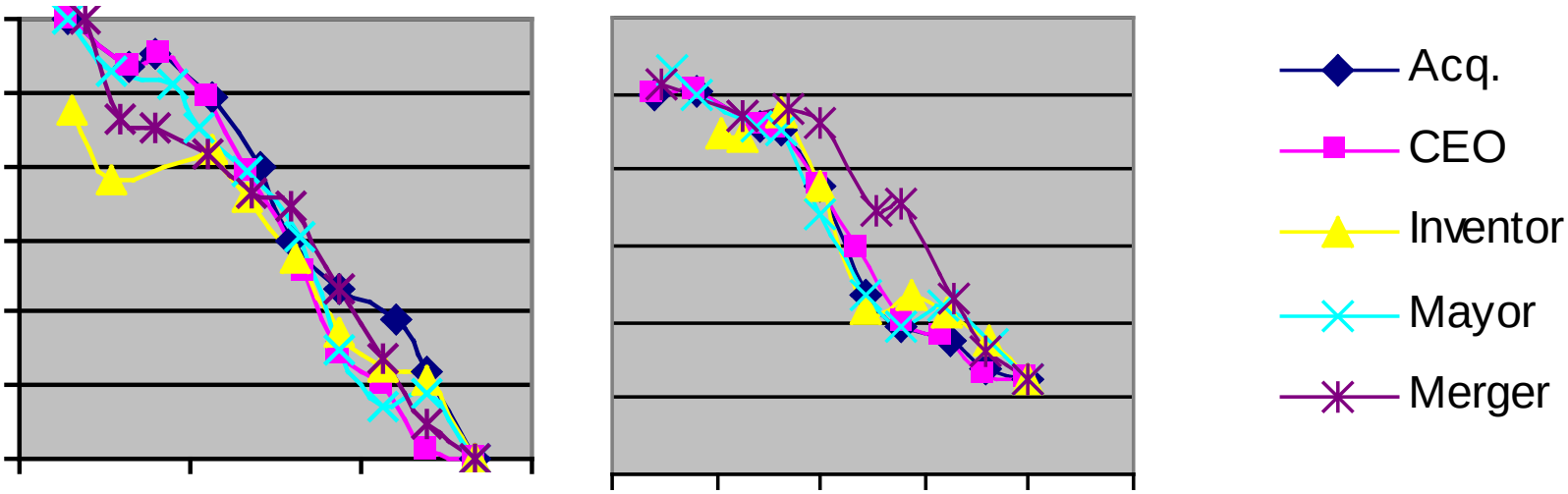
Experimental Evaluation

- We want to answer the following questions:
  1. Can we train SRES's classifier once, and then use the results on all other relations?
  2. How does SRES's performance compare with KnowItAll and KnowItAll-PL?

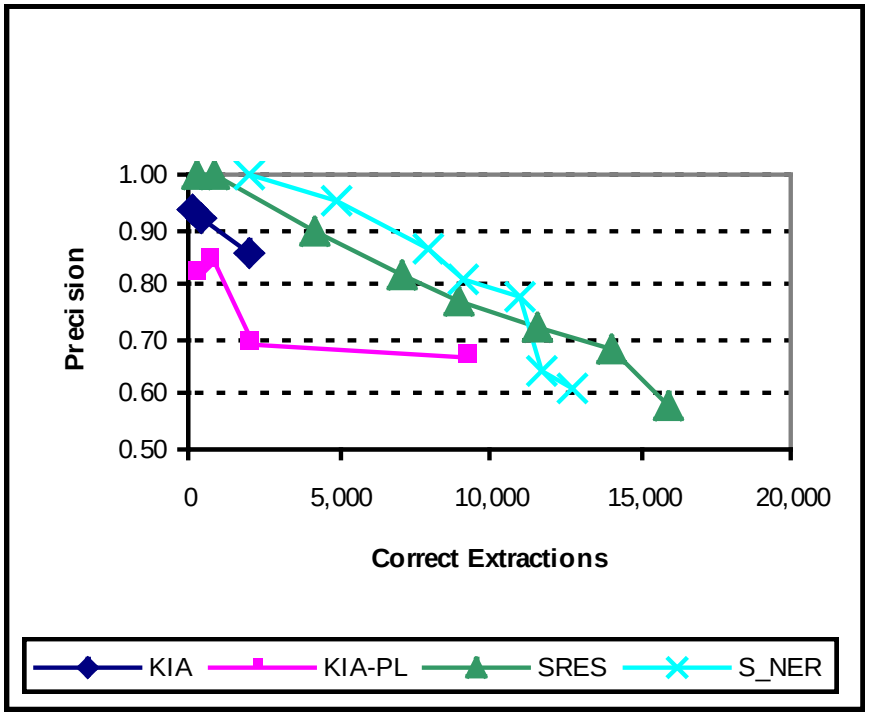
Sample Output

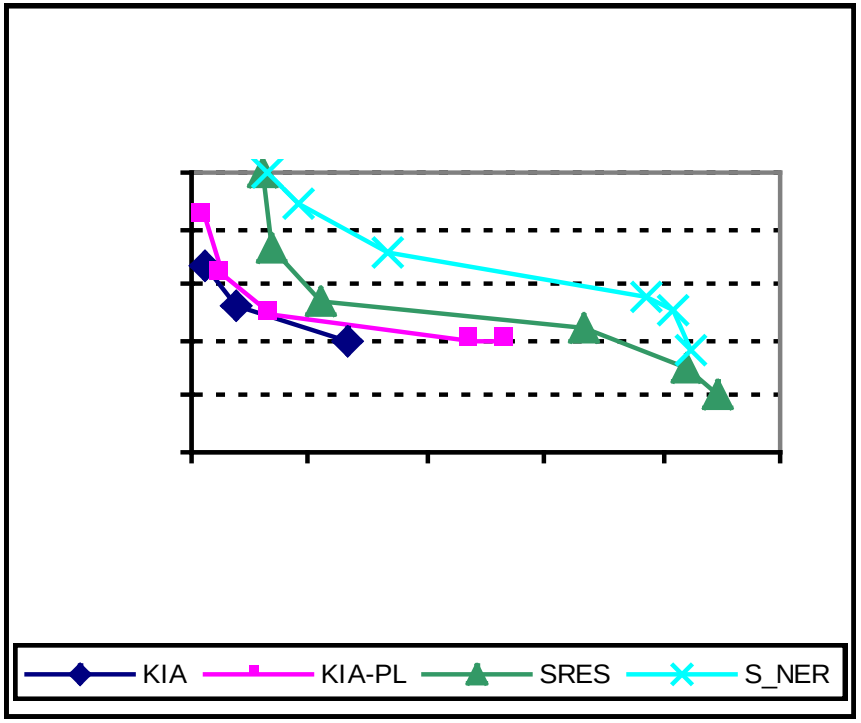
- **HP – Compaq merger**
  - <s><DOCUMENT>Additional information about the <X>HP</X> -<Y>Compaq</Y> merger is available at www.VotetheHPway.com .</DOCUMENT></s>
  - <s><DOCUMENT>The Packard Foundation, which holds around ten per cent of <X>HP</X> stock, has decided to vote against the proposed merger with <Y>Compaq</Y>.</DOCUMENT></s>
  - <s><DOCUMENT>Although the merger of <X>HP</X> and <Y>Compaq</Y> has been approved, there are no indications yet of the plans of HP regarding Digital GlobalSoft.</DOCUMENT></s>
  - <s><DOCUMENT>During the Proxy Working Group's subsequent discussion, the CIO informed the members that he believed that Deutsche Bank was one of <X>HP</X>'s advisers on the proposed merger with <Y>Compaq</Y>.</DOCUMENT></s>
  - <s><DOCUMENT>It was the first report combining both <X>HP</X> and <Y>Compaq</Y> results since their merger.</DOCUMENT></s>
  - <s><DOCUMENT>As executive vice president, merger integration, Jeff played a key role in integrating the operations, financials and cultures of <X>HP</X> and <Y>Compaq</Y> Computer Corporation following the 19 billion merger of the two companies.</DOCUMENT></s>

Cross-Classification Experiment

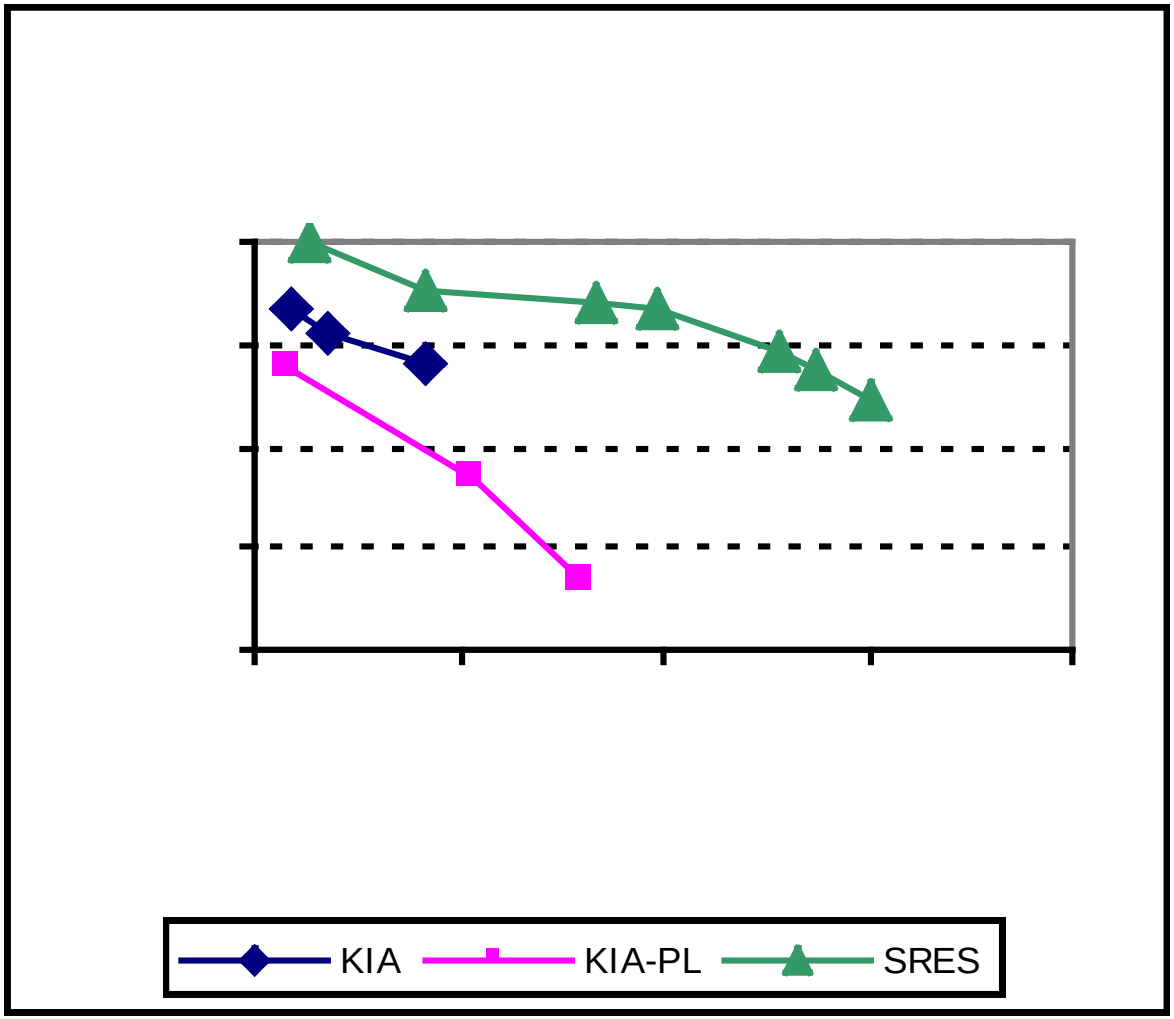


Results!


1.00
0.90
0.80
0.70
0.60
0.50
Precision
0
5,000
10,000
15,000
20,000
Correct Extractions
KIA
KIA-PL
SRES
S_NER


KIA
KIA-PL
SRES
S_NER

Inventor Results

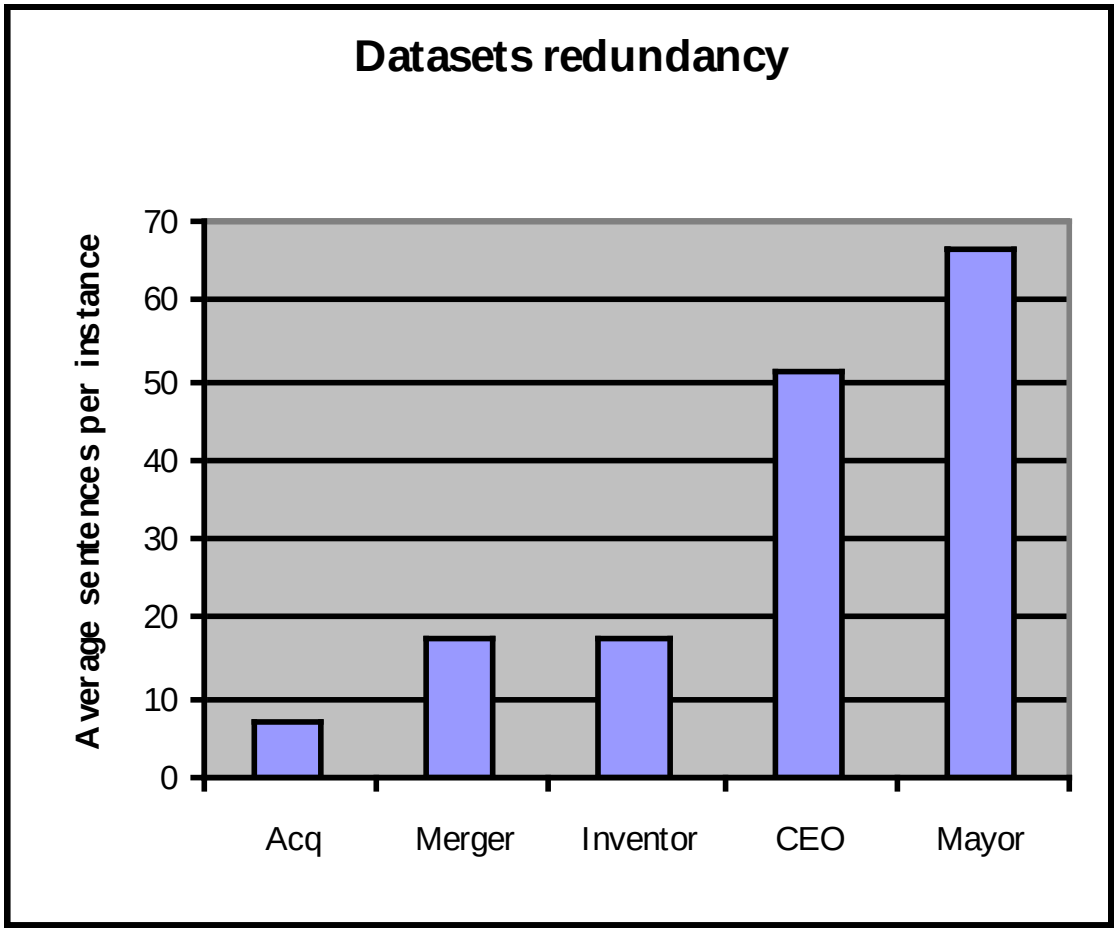


When is SRES better than KIA?

- KnowItAll extraction works well when
  - redundancy is high
  - most instances have a good chance of appearing in simple forms

- SRES is more effective for low-frequency instances due to
  - more expressive rules
  - classifier that inhibits those rules from overgeneralizing.

The Redundancy of the Various Datasets



Outline

- Intro to text mining
- Information extraction (IE)
- Open IE/Relation Extraction
- Sentiment mining
  - And its relation to IE: CARE
- Wrap-up

Advanced Approaches to information extraction

Ronen Feldman
Hebrew University
Jerusalem, israel
Ronen.feldman@huji.ac.il

Traditional Text Mining is not cost effective nor time efficient

- Why?
  - takes too long to develop
  - too expensive
  - not accurate enough
  - lacks complete coverage

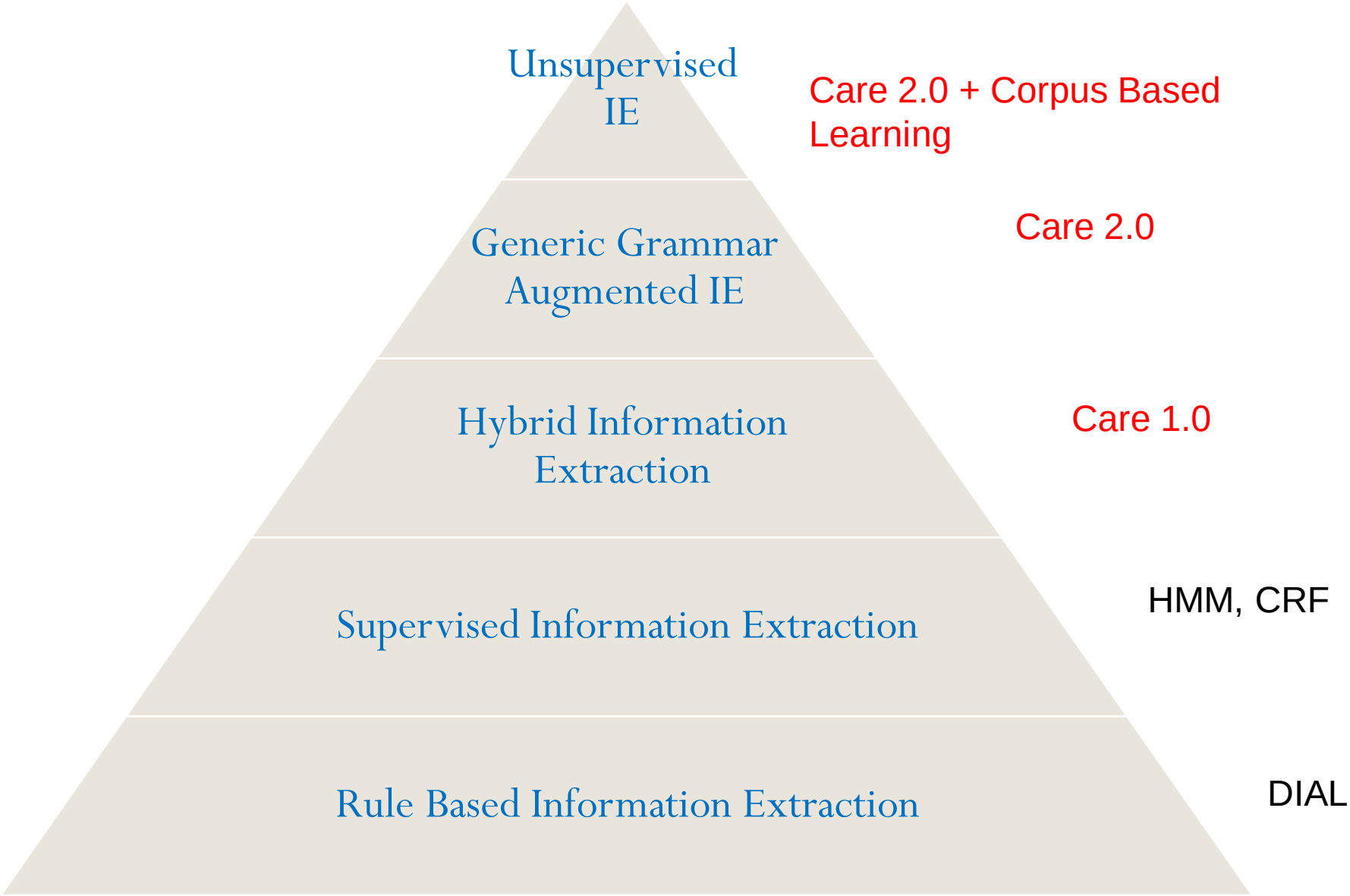
A pure statistical approach is not accurate enough

- Without semantic comprehension - is Apple a fruit or a company?
- 'We reduced our deficit' – need proper negation handling

Rule writing approach

- Domain specific
- Very long development cycle
- Expensive process
- No guarantee of full pattern coverage

The Evolution of Information Extraction Technology



Example of Unsupervised IE Results

Actos; weight gain (40 (P: 38, N: 2))

Rel_take_DRU_has_SYM(DRUG, SYMPTOM)

Negative (1)

I've been taking 15 mg of **Actos** for just over a year now and so far (knock on wood) I haven't had the weight gain that
some others have reported as a side effect.

Positive (8)

I also have read here about some of you who have been on the **Actos** and the weight gain you had experienced.
We saw an endo because of all of the weight gain and side effects from taking **actos**.
He was on **Actos** but went off of it because of weight gain and stomach bloating.
I really don't want to go back on **Actos** because of weight gain/fluid retention.
My doctor wanted me to start **Actos** for awhile, until the Byetta kicks in, but I stopped **Actos** in the first place because
of weight gain and I said no to restarting that.
I started taking **Actos** first on May 2, 2007 and I started Metformin 3 weeks later I can not take the Spironolactone till
Aug but I have noticed that I have gained weight with these 2 drugs instead of losing
and I got a treadmill and do 30 min every morning when I get up and lately I have been doing 30 min at night too
because of the weight gain.
I have experienced weight gain as well and i am on **Actos** and insulin and glucophage.
I guess that everything comes with a price, but I'm wondering if most folks who have tried **Actos** have experienced
weight gain and the other side effects (edema, headaches, nausea, fatigue, etc.).

Rel_SYMofDRU(SYMPTOM, DRUG)

Positive (5)

I do notice that it increases my hunger, so it is possible that **Actos** weight gain issues may be from hunger being
stimulated.
I don't think that a lot of us had made the **Actos** induced weight gain connection.
One reported side effect of **Actos** is weight gain.
I have changed to a new MD and when I discussed my concern over the weight gain with Avandia and then
Actos, he

Example of Unsupervised IE Results

Rel_cause_DRUvSYM(DRUG, SYMPTOM)

Negative (1)

Actos hasn't caused any weight gain, I am still losing some.

Positive (25)

I also am on Synthroid, Atenolol, Diovan, Lotrel, Lexapro, Vitorin and Prilosec OTC. I didn't realize that **Actos** can cause a weight gain

as I had never read it as a side effect; however, after reading all of the

comments on this site, I now know why my weight has increased over the past few months since taking on it.

I don't take any oral meds, but from what I have read here, **Actos** causes weight gain because of water retention.

why does the endo think you're a type 1? oral meds are usually given only to type 2's,as type 2's have insulin resistance. oral meds

treat the insulin resistance. type 1's require insulin.....i take actoplus met- which is actos and metformin.actos is like avandia and

i've had no heart issues.....tho-avandia and **actos** can cause weight gain....take care,trish

Actos causes edema and weight gain also.

Actos can cause weight gain (so can Avandia, it's cousin)

Now I have started to see a lot of reports of **Actos** causing weight gain, among other things.

for the record, **Actos** can, and does, cause weight gain/water retention.

I'm on both - what did you hate about Metformin? (**Actos** causes weight gain, metformin weight loss)

Also I hear that the **Actos** causes weight gain, so now I am afraid the new pill will cause me to gain weight.

I'm type 1 so only on insulin, but I have heard that **Actos** can cause weight gain.

Avandia & **Actos**, especially in combination with insulin, causes fluid retention and/or fat weight gain.

My endocrinologist warned me that **Actos** can cause significant weight gain.

Actos caused weight gain and fluid retention in my chest.

Metformin causes weight loss, Avandia and **Actos** causes the birth of new fat cells and weight gain.

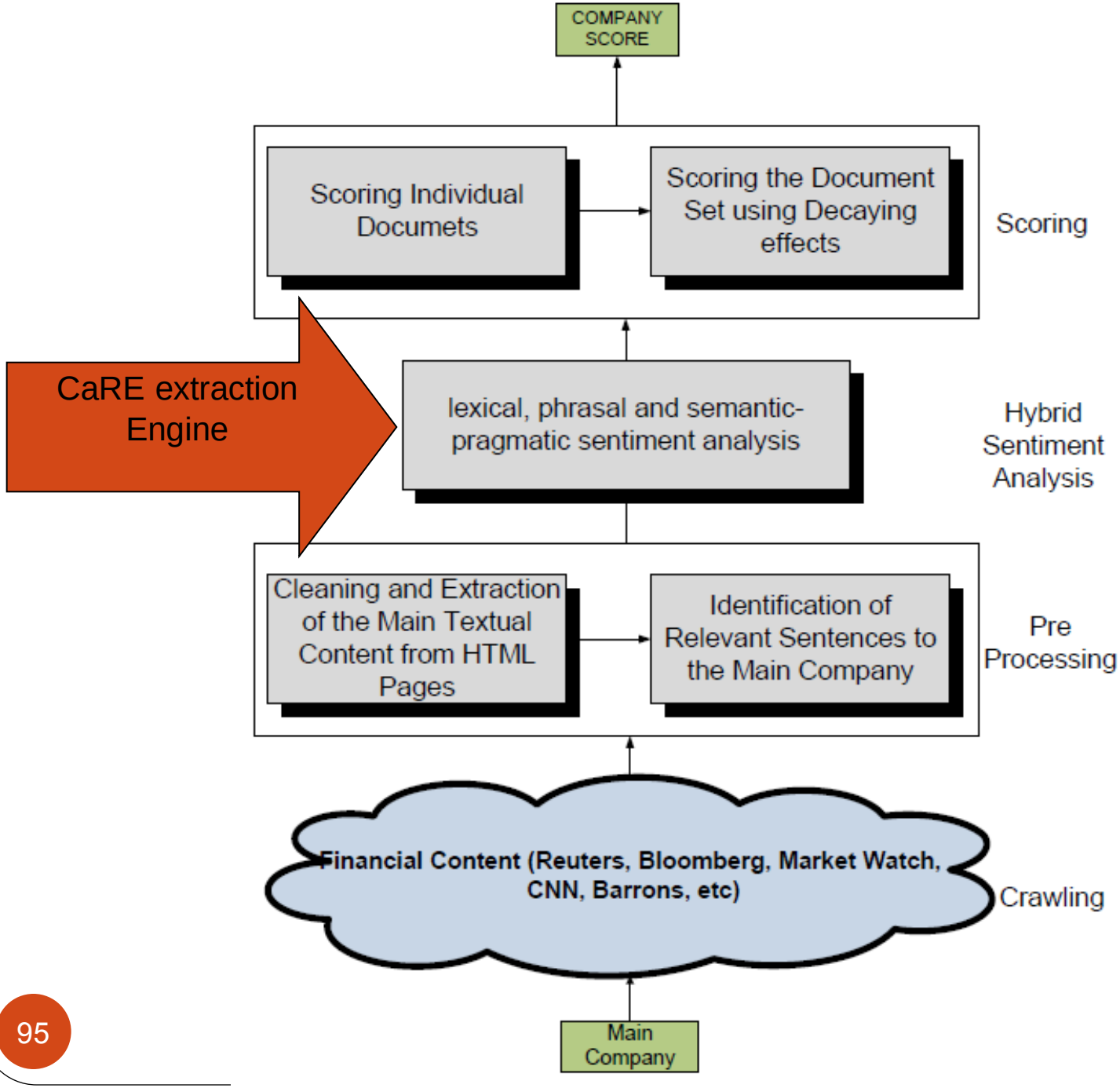
……

Sentiment Analysis of Stocks from News Sites

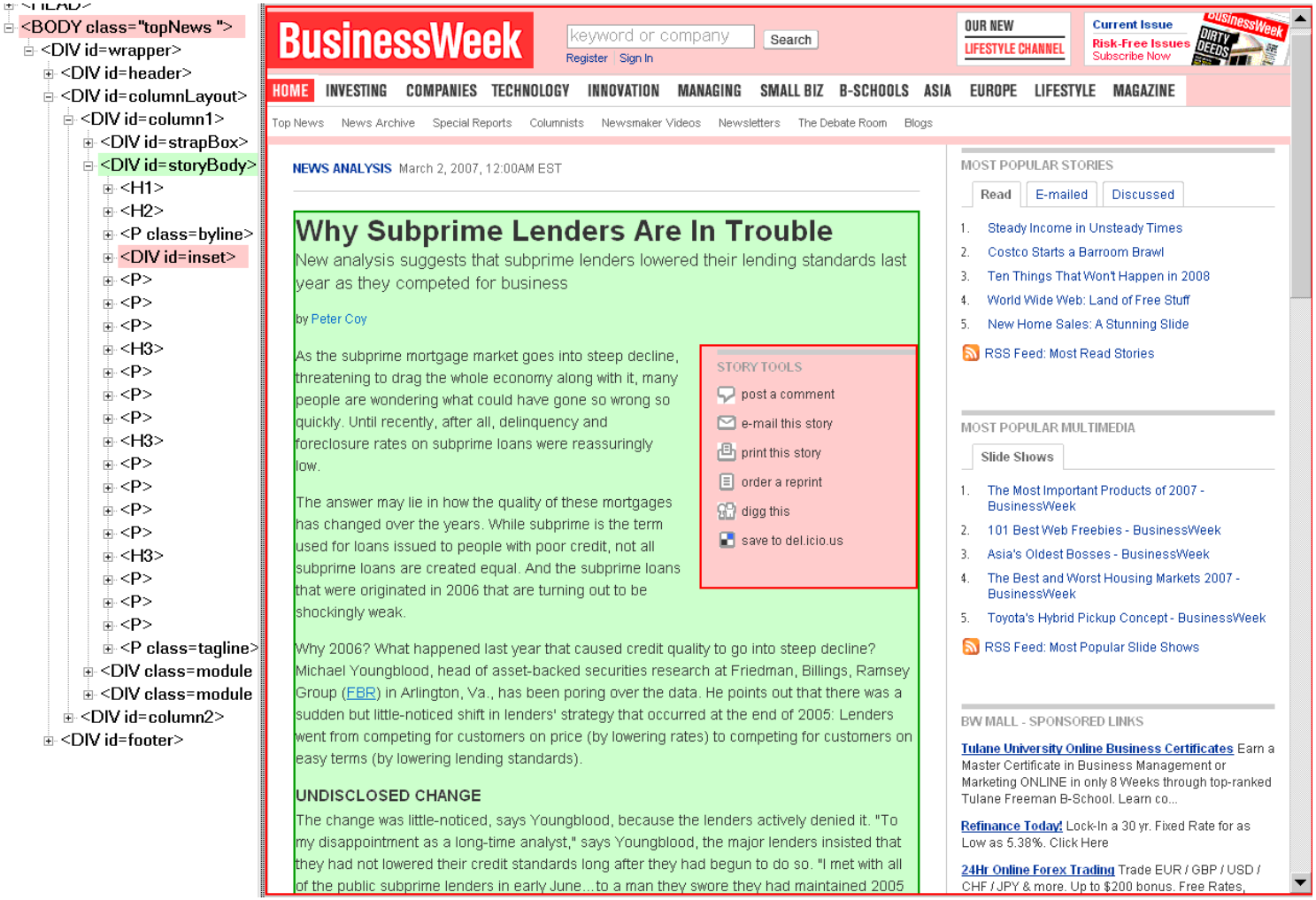
The Need for Event Based SA

*Toyota announces voluntary recall of their **highly successful top selling** 2010 model-year cars*

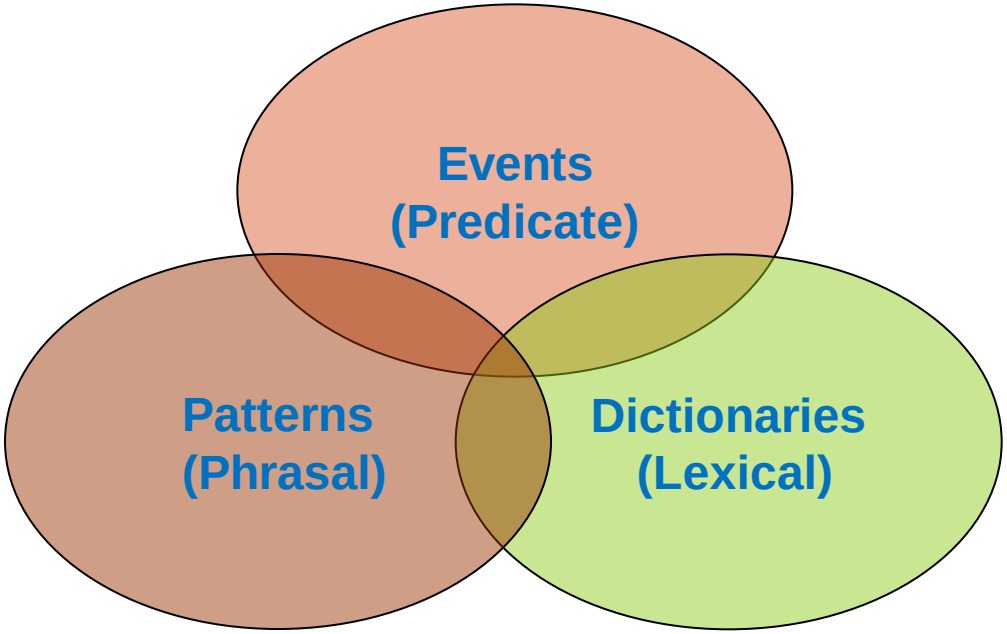
- Phrase-level SA:
  - highly successful top selling ⇒ **positive**
  - Or at best neutral
    - Taking into account voluntary recall ⇒ **negative**
- Need to recognize the whole sentence as a "product recall" event!


COMPANY SCORE
Scoring Individual Documets
Scoring the Document Set using Decaying effects
Scoring
CaRE extraction Engine
lexical, phrasal and semantic-pragmatic sentiment analysis
Hybrid Sentiment Analysis
Cleaning and Extraction of the Main Textual Content from HTML Pages
Identification of Relevant Sentences to the Main Company
Pre Processing
Financial Content (Reuters, Bloomberg, Market Watch, CNN, Barrons, etc)
Crawling
Main Company

Template Based Approach to Content Filtering



Hybrid Sentiment Analysis



- All levels are part of the same rulebook, and are therefore considered simultaneously by CaRE

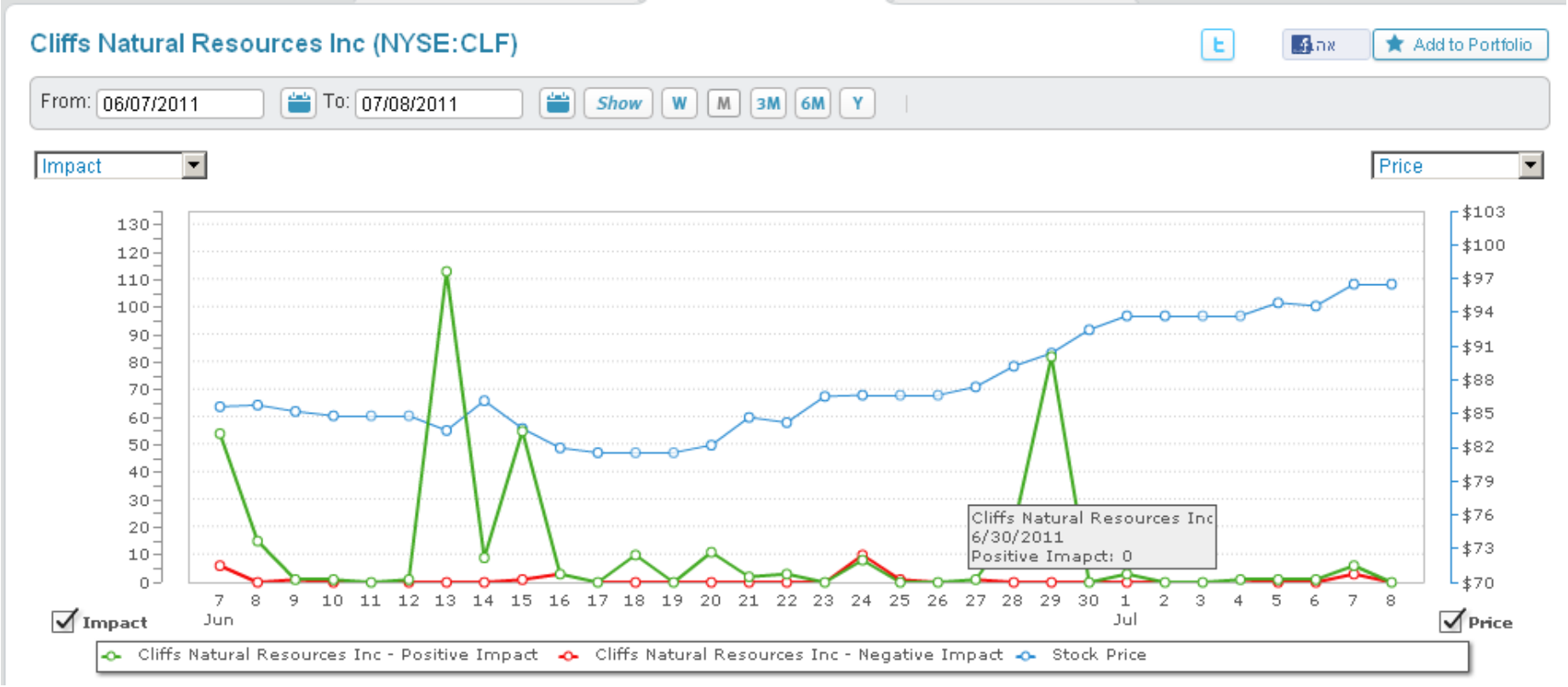
Dictionary-based sentiment

- Started with available sentiment lexicons
  - Domain-specific and general
  - Improved by our content experts
- Examples
  - Modifiers: attractive, superior, inefficient, risky
  - Verbs: invents, advancing, failed, lost
  - Nouns: opportunity, success, weakness, crisis
  - Expressions: exceeding expectations, chapter 11
- Emphasis and reversal
  - successful, **extremely successful,**
    far from successful

Event-Based Sentiment

- Product release/approval/recall, litigations, acquisitions, workforce change, analyst recommendations and many more
- Semantic role matters:
  - Google is being sued/is suing…
- Need to address historical/speculative events
  - Google acquired YouTube **in 2006**
  - **What if Google buys Yahoo** and the software giant Microsoft remains a single company fighting for the power of the Internet?

CLF


Cliffs Natural Resources Inc (NYSE:CLF)
Add to Portfolio
From: 06/07/2011
To: 07/08/2011
Show
W
M
3M
6M
Y
Impact
Price
130
120
110
100
90
80
70
60
50
40
30
20
10
0
$103
$100
$97
$94
$91
$88
$85
$82
$79
$76
$73
$70
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 1 2 3 4 5 6 7 8
Jun
Jul
Cliffs Natural Resources Inc
6/30/2011
Positive Imapct: 0
Impact
Price
Cliffs Natural Resources Inc - Positive Impact
Cliffs Natural Resources Inc - Negative Impact
Stock Price

Why did we get a Positive Spike?

Articles | Events

Filter

By Impact:

Negative | Neutral | Positive

Show neutral articles

Set Filter

Sort by: Most Recent

+0.76 [Neg:0 | Pos:10]	Cliffs Natural Resources Gets Into Dilution 6/13/2011 | forbes.com
+0.40 [Neg:0 | Pos:2]	(AMM) Analysts upbeat on Cliffs growth, offering of shares 6/13/2011 | metalbulletin.com
+0.93 [Neg:0 | Pos:41]	UBS Initiates Coverage of Cliffs Natural Resources With a Buy Rating and A $122 Target Price 6/13/2011 | fnno.com
+0.86 [Neg:0 | Pos:20]	UBS Initiates Coverage on Cliffs Natural Resources 6/13/2011 | streetinsider.com
+0.76 [Neg:0 | Pos:10]	On The Fly: Analyst Initiation Summary 6/13/2011 | finance.yahoo.com
+0.76 [Neg:0 | Pos:10]	Nokia, LinkedIn: Analysts' New Ratings 6/13/2011 | thestreet.com
+0.86 [Neg:0 | Pos:20]	UBS Starts Cliffs Natural Resources 6/13/2011 | streetinsider.com

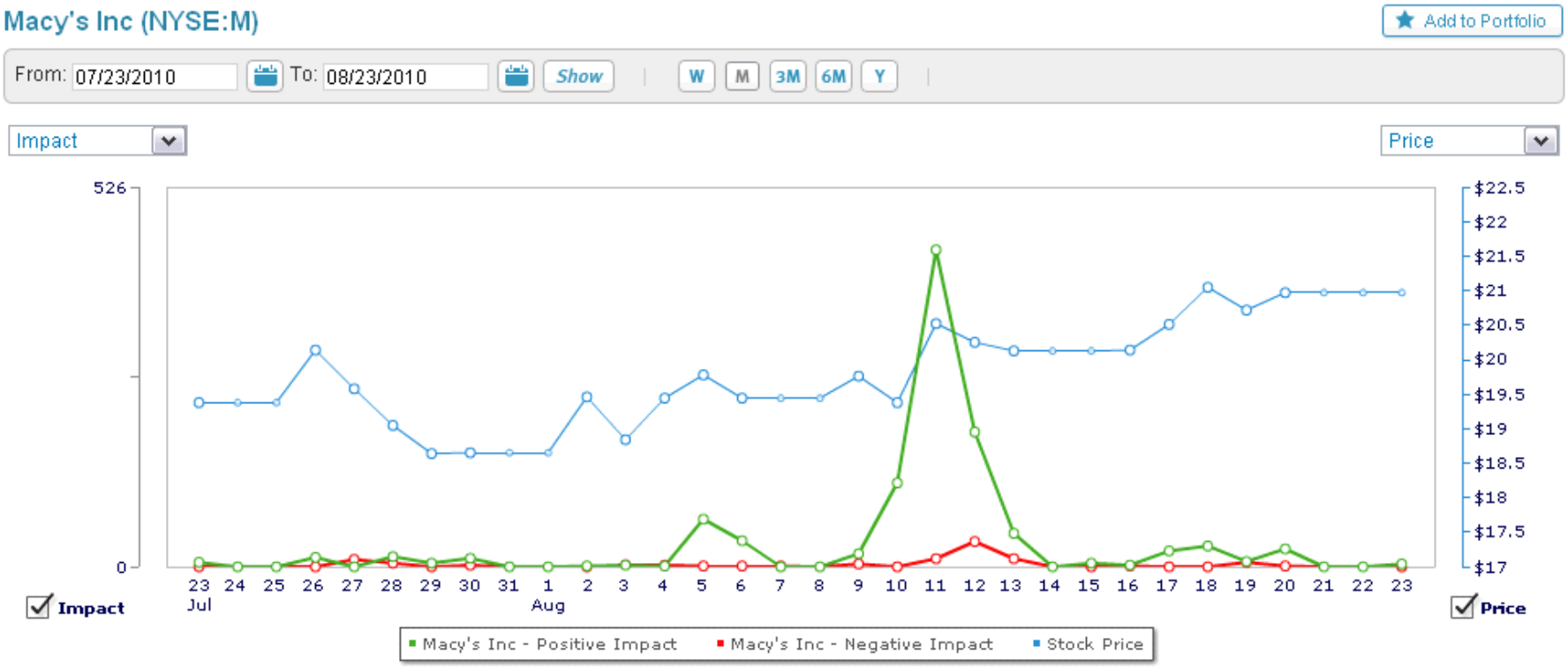
Results 1-7 of 60

Article Preview

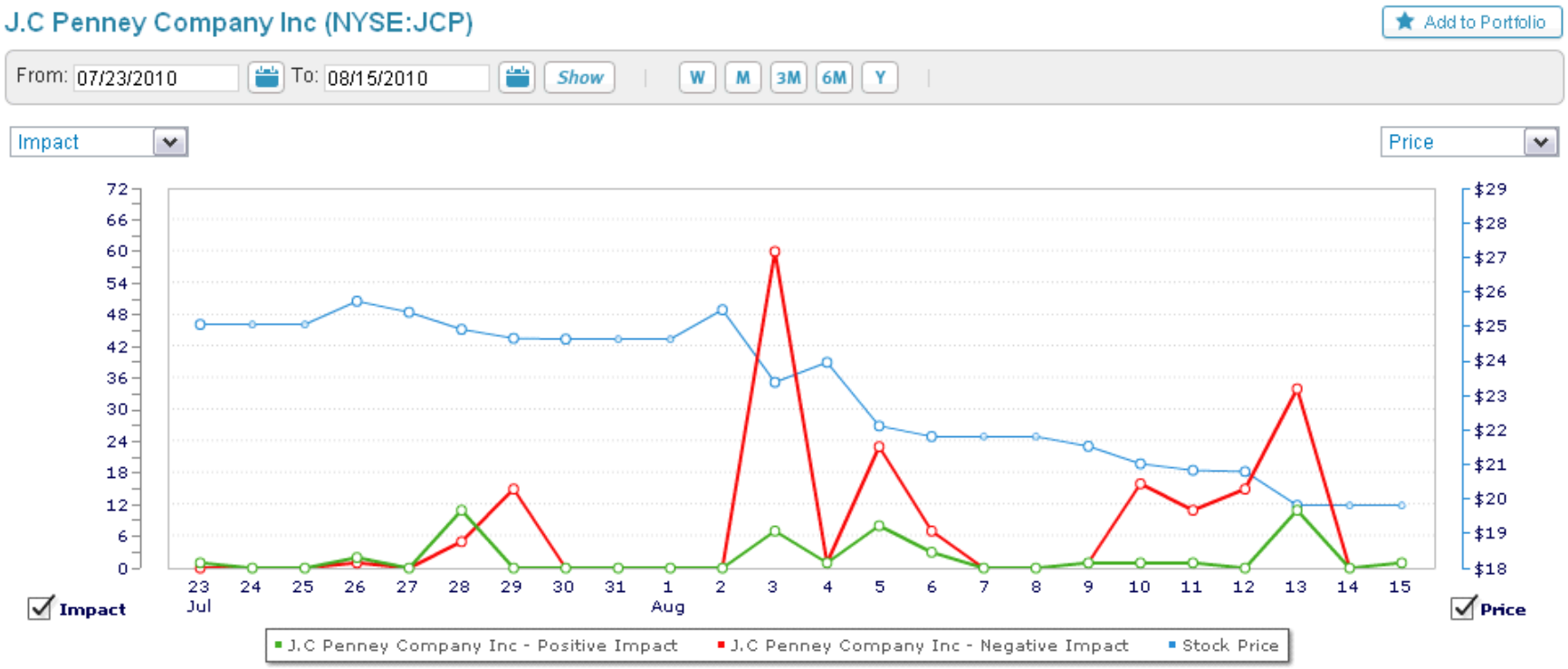
UBS Initiates Coverage of Cliffs Natural Resources With a Buy Rating and A $122 Target Price

UBS initiated coverage of Cliffs Natural Resources (NYSE:CLF) today with a buy rating and a $122 price target. The bank says the company's seaborne exposure drives strong projected cash flow, as the supply for seaborne iron ore will be tight over the next two years. The Consolidated Thompson acquisition increases Cliffs' seaborne exposure and diversifies the company's exposure to U.S. iron ore and coal segments, according to UBS. Cliffs Natural Resources has a potential upside of 50.3% based on a current price of $85.04 and an average consensus analyst price target of $127.83. By Tim Tracy ttracy@fnno.com Latest Breaking Video News Morgan Stanley lifted its price target for Air Products and Chemicals (NYSE:APD) to $115 from $105 and maintained its overweight rating on the stock. The bank sees 2011 EPS of $5.90, up from $5.76 per share, and 2012 EPS of $6.63, up from $6.45 per share. Air Products &... Citigroup maintained its buy rating on the New York Times (NYSE:NYT), with an $11.50 price target, but said the key risk at this point is that core newspaper advertising revenues remain softer than expected. The bank sees 2011 EPS of $0.54 and 2012 EPS of $0.87. The New York Times...

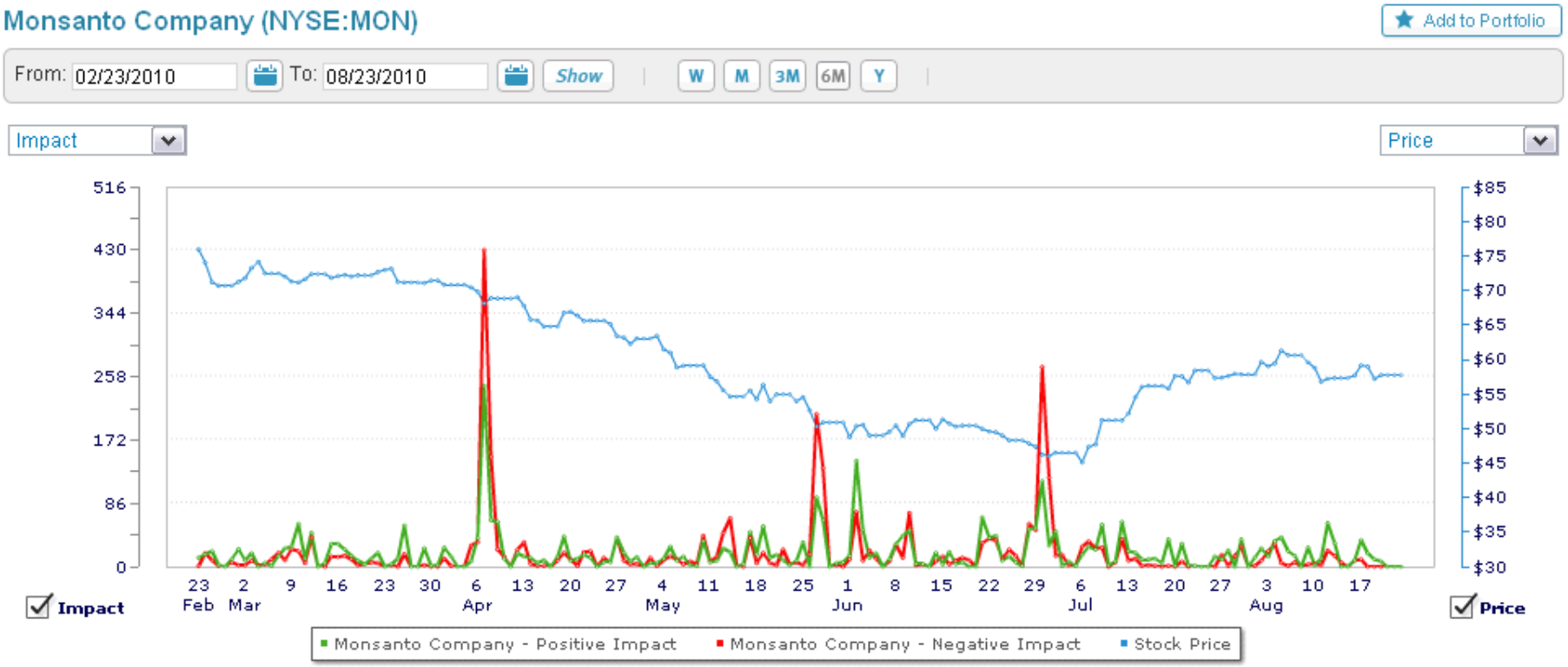
Macy's


Macy's Inc (NYSE:M)
Add to Portfolio
From: 07/23/2010
To: 08/23/2010
Show
W
M
3M
6M
Y
Impact
Price
526
0
$22.5
$22
$21.5
$21
$20.5
$20
$19.5
$19
$18.5
$18
$17.5
$17
23 24 25 26 27 28 29 30 31 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Jul
Aug
Impact
Price
Macy's Inc - Positive Impact
Macy's Inc - Negative Impact
Stock Price

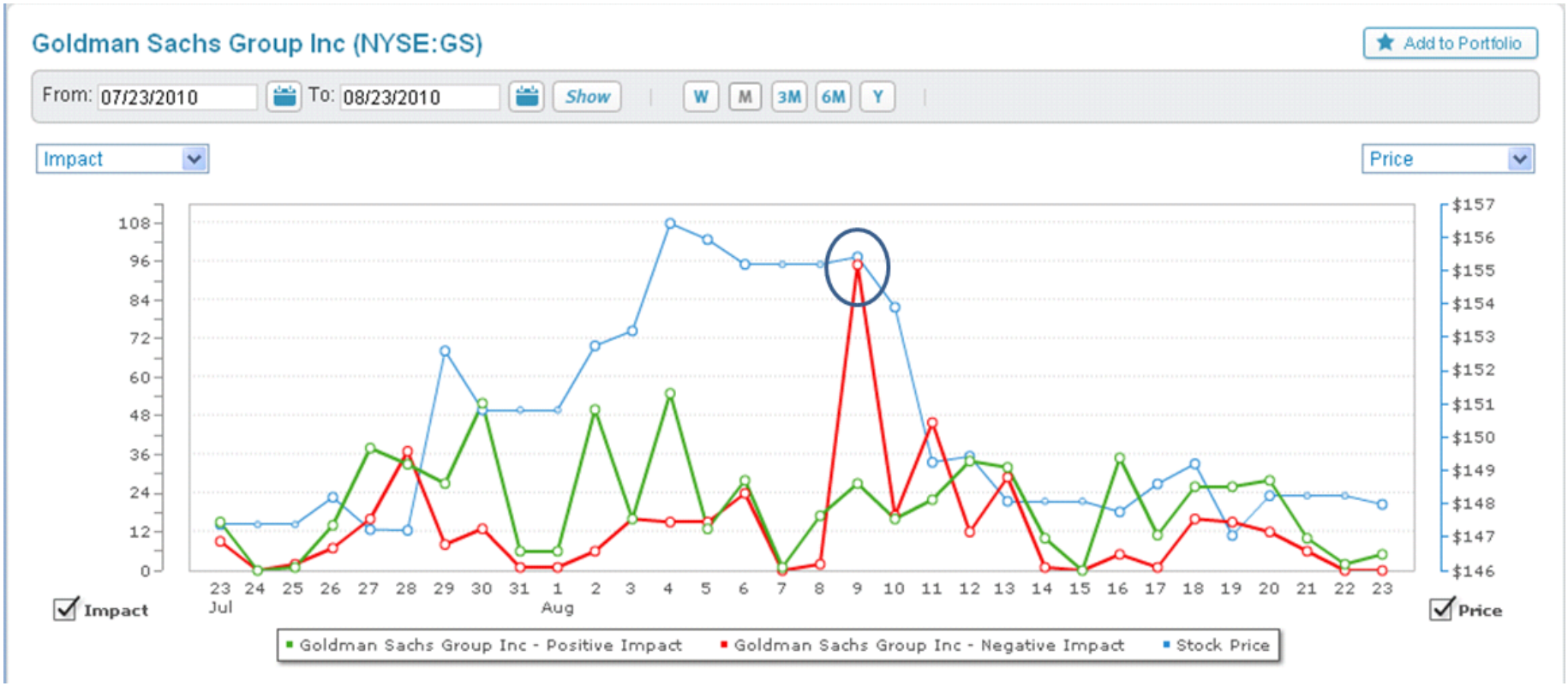
JC Penny


J.C Penney Company Inc (NYSE:JCP)
Add to Portfolio
From: 07/23/2010
To: 08/15/2010
Show
W
M
3M
6M
Y
Impact
Price
72
66
60
54
48
42
36
30
24
18
12
6
0
$29
$28
$27
$26
$25
$24
$23
$22
$21
$20
$19
$18
23 24 25 26 27 28 29 30 31 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Jul
Aug
Impact
Price
J.C Penney Company Inc - Positive Impact
J.C Penney Company Inc - Negative Impact
Stock Price

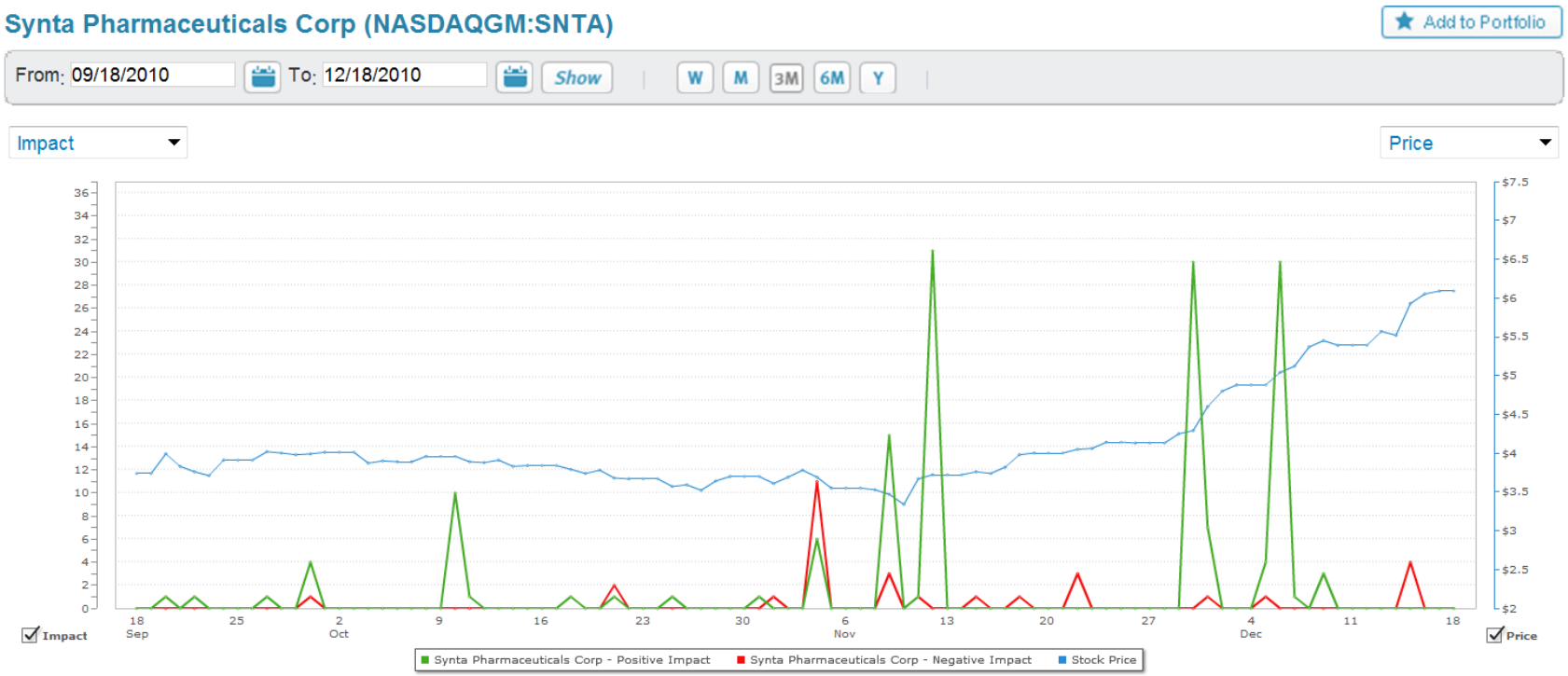
Monsanto


Monsanto Company (NYSE:MON)
Add to Portfolio
From: 02/23/2010
To: 08/23/2010
Show
W
M
3M
6M
Y
Impact
Price
516
430
344
258
172
86
0
$85
$80
$75
$70
$65
$60
$55
$50
$45
$40
$35
$30
23 Feb
2 Mar
9
16
23
30
6 Apr
13
20
27
4 May
11
18
25
1 Jun
8
15
22
29
6 Jul
13
20
27
3 Aug
10
17
Impact
Price
Monsanto Company - Positive Impact
Monsanto Company - Negative Impact
Stock Price

Goldman Sachs



SNTA 3 Months



Key Developments

Key Developments - Windows Internet Explorer

http://keydevelopments.digitaltrowel.com/

File Edit View Favorites Tools Help

Favorites | Key Developments | Page | Safety | Tools

Key Developments Prototype

digital trowel
Pure Intelligence.

Companies

- The Boeing Company
- Textron Inc
- Pfizer Inc
- Cisco Systems Inc
- Goldman Sachs Group Inc
- Exxon Mobil Corporation
- Ford Motor Company
- Starwood Hotel & Res...
- Covance Inc

Headlines

Min. Significance Score: 0.8

Boeing, National Renewable Energy Lab to Collaborate on DOD Energy Security

Article Date: 10/26/2010
Significance Score:

Boeing, National Renewable Energy Lab to Collaborate on DOD Energy Security

Boeing Receives Secure Messaging Technology Contract Extension from US Army

Article Date: 10/18/2010
Significance Score: 0.84

Boeing and Northrop Grumman GMD Team Announces Partners for Competitive Missile Defense Contract

Article Date: 10/12/2010
Significance Score: 0.86

Boeing Receives Contract for Apache Helicopter Rotor Heads

Article Date: 10/4/2010
Significance Score: 0.85

Boeing Receives A-10 Structural Inspection Contract from US Air Force

Article Date: 9/29/2010
Significance Score: 0.87

Boeing Awarded $19 Million Follow-on US Navy Field Services Contract

Key Developments

Boeing Receives Contract for Apache Helicopter Rotor Heads

Boeing [NYSE: BA] today announced that the U.S. Defense Logistics Agency s awarded the company a three-year contract to manufacture rotor heads for the U.S. Army's fleet of Apache attack helicopters. The fixed-price contract calls for an initial 100 main rotor heads, at a value of $31 million, and allows for 50 additional rotor heads, which would increase the contract value to $48 million.

Full Article

Boeing Receives Contract for Apache Helicopter Rotor Heads

10/4/2010 | The Boeing Company corporate web site

MESA, Ariz., Oct. 4, 2010 -- **Boeing [NYSE: BA] today announced that the U.S. Defense Logistics Agency Aviation has awarded the company a three-year contract to manufacture spare main rotor heads for the U.S. Army's fleet of Apache attack helicopters. The fixed-price contract calls for an initial 100 main rotor heads, at a value of $31 million, and allows for 50 additional rotor heads, which would increase the contract value to $48 million.**

The Army Integrated Logistics division of Boeing Global Services & Support will perform the work at the Boeing facility in Mesa. "This contract award is another testament to Boeing employees' excellent work and

Mining Medical User Forums

The Text Mining Process

Downloading
- html-pages are downloaded from a given forum site

Cleaning
- html-like tags and non-textual information like images, commercials, etc… are cleaned from the downloaded text

Chunking
- The textual parts are divided into informative units like threads, messages, and sentences

Information Extraction
- Products and product attributes are extracted from the messages

Comparisons
- Comparisons are made either by using co-occurrence analysis or by utilizing learned comparison patterns

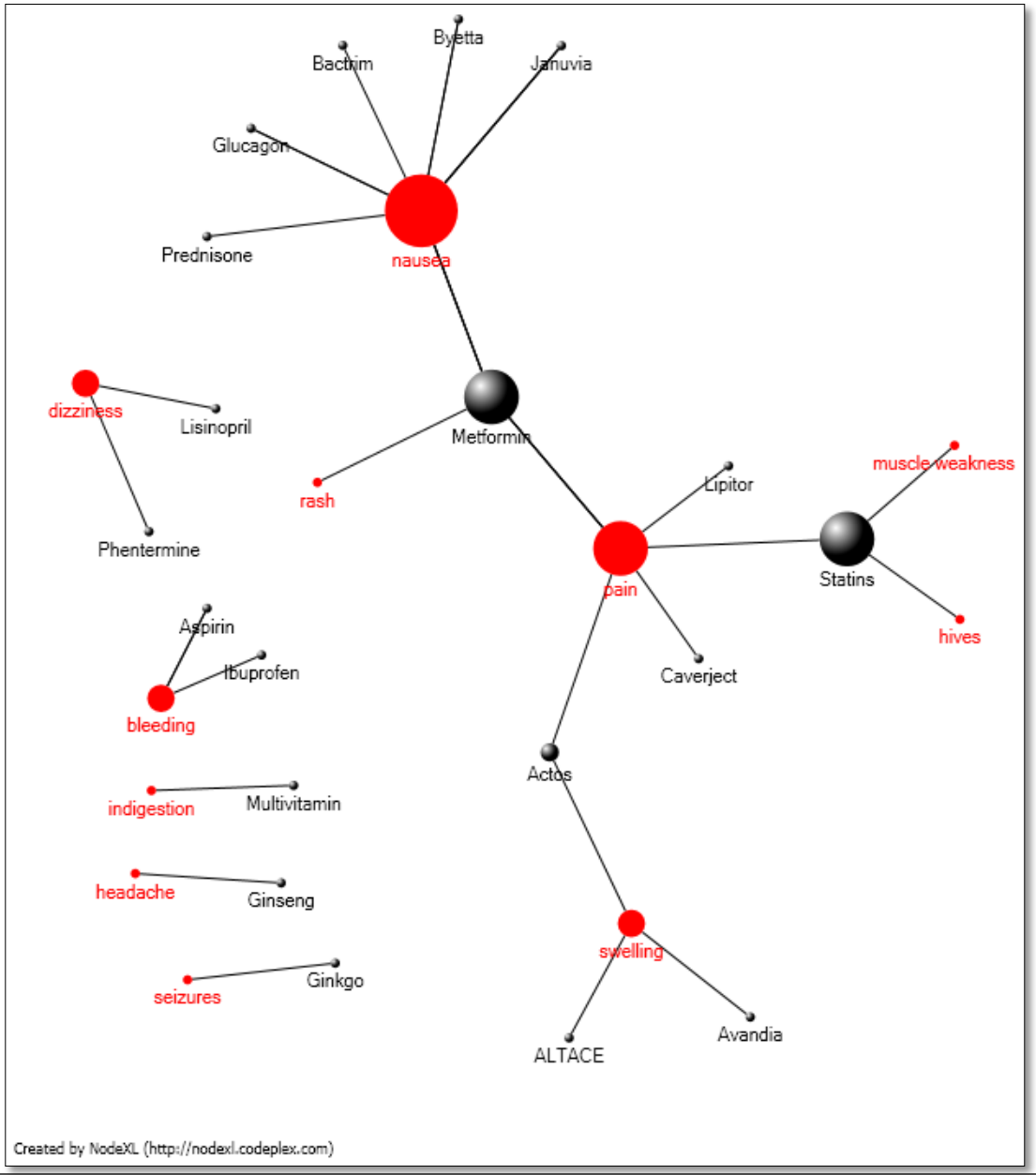
The Text Mining Process

We downloaded messages from 5 different consumer forums

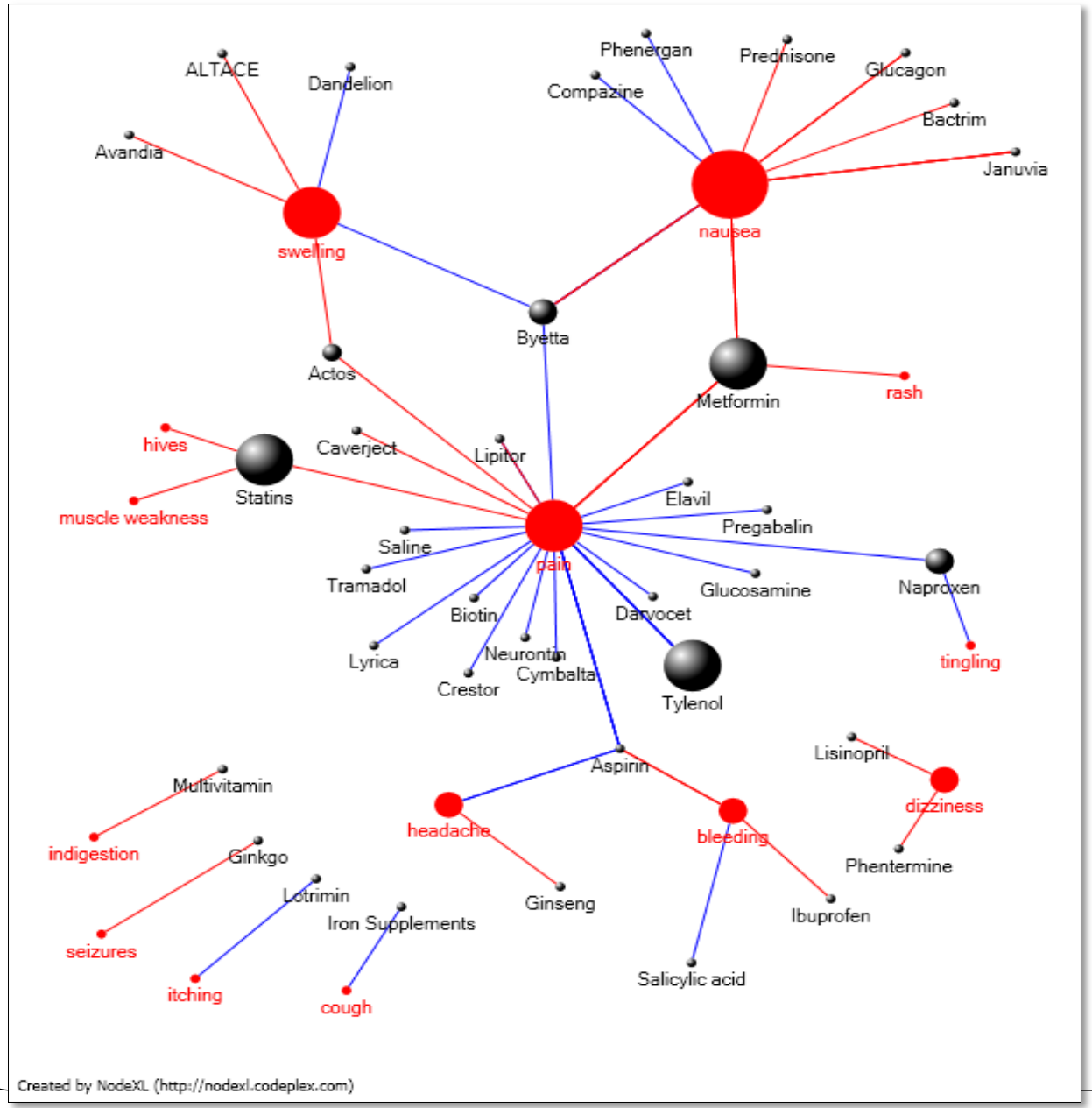
- diabetesforums.com
- healthboards.com
- forum.lowcarber.org
- diabetes.blog.com**
- diabetesdaily.com

** Messages in Diabets.blog.com were focused mainly on Byetta

Side Effects



Side Effects and Remedies



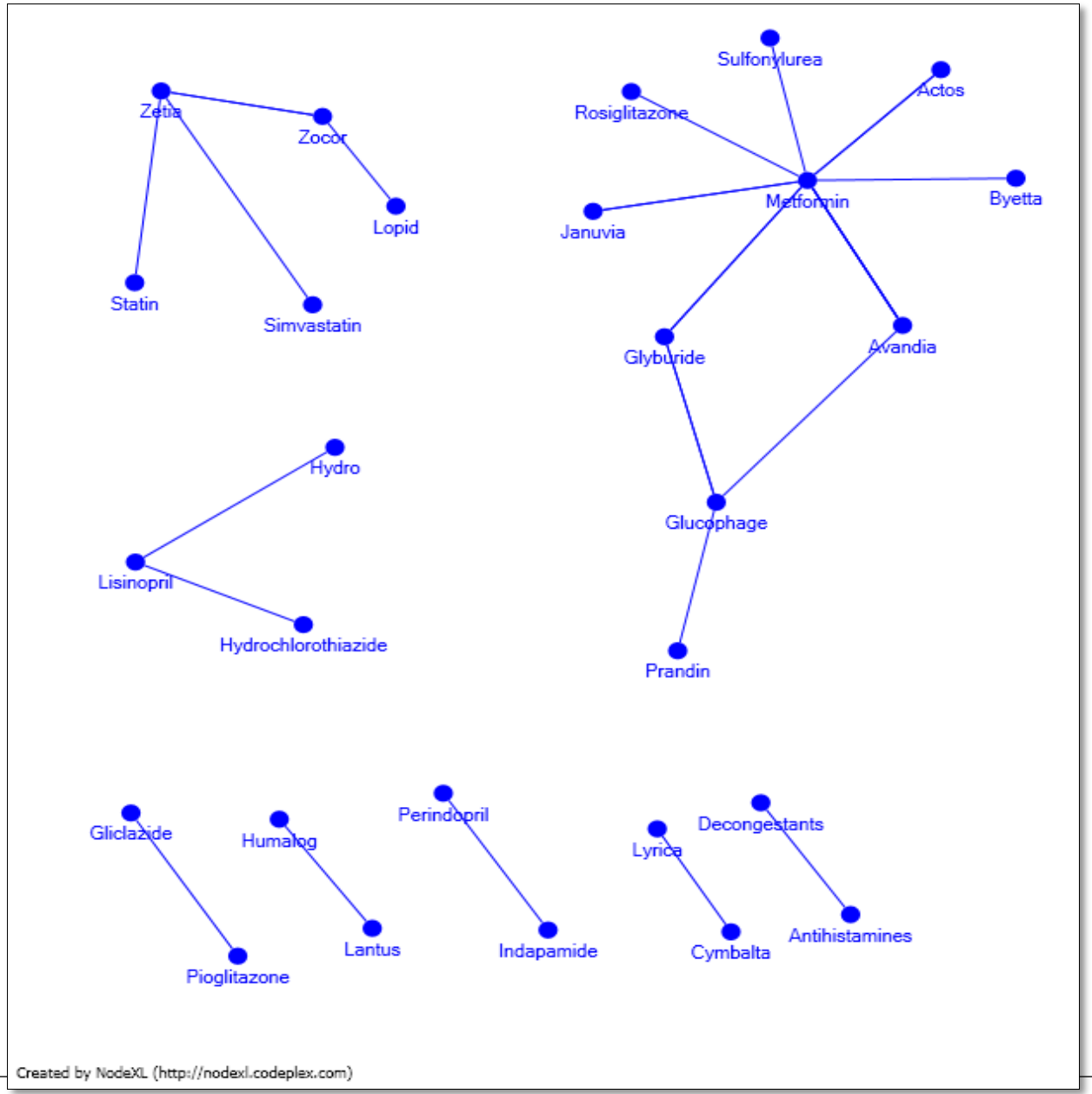
Red lines – side effects/symptoms
Blue lines - Remedies

See what causes symptoms and what relieves them

See what positive and negative effects a drug has

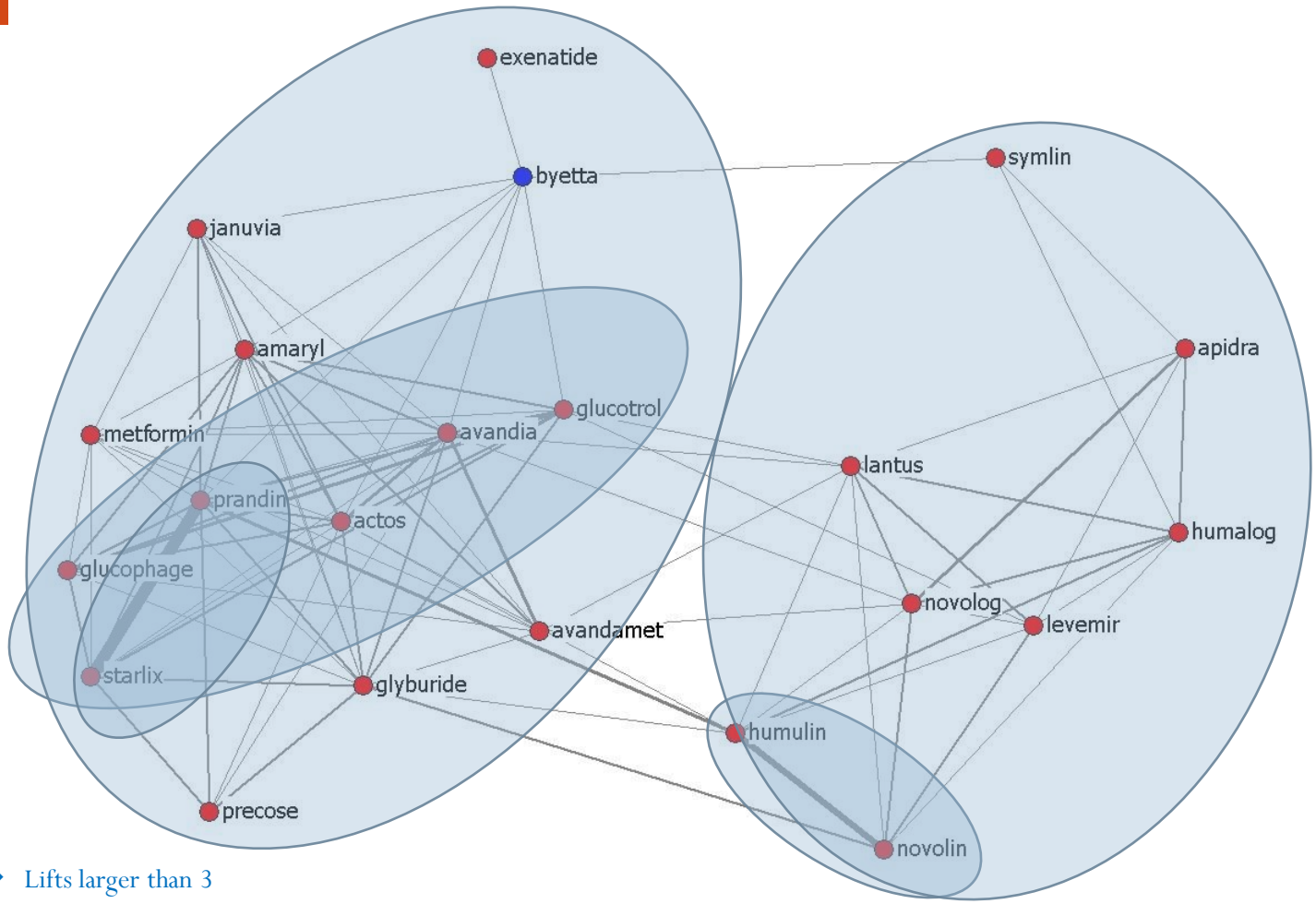
See which symptoms are most complained about

Drugs Taken in Combination



Drug Analysis

Drug Co-Occurrence - Spring Graph – Perceptual Map



Several Pockets of drugs that were mentioned frequently together in a message were identified

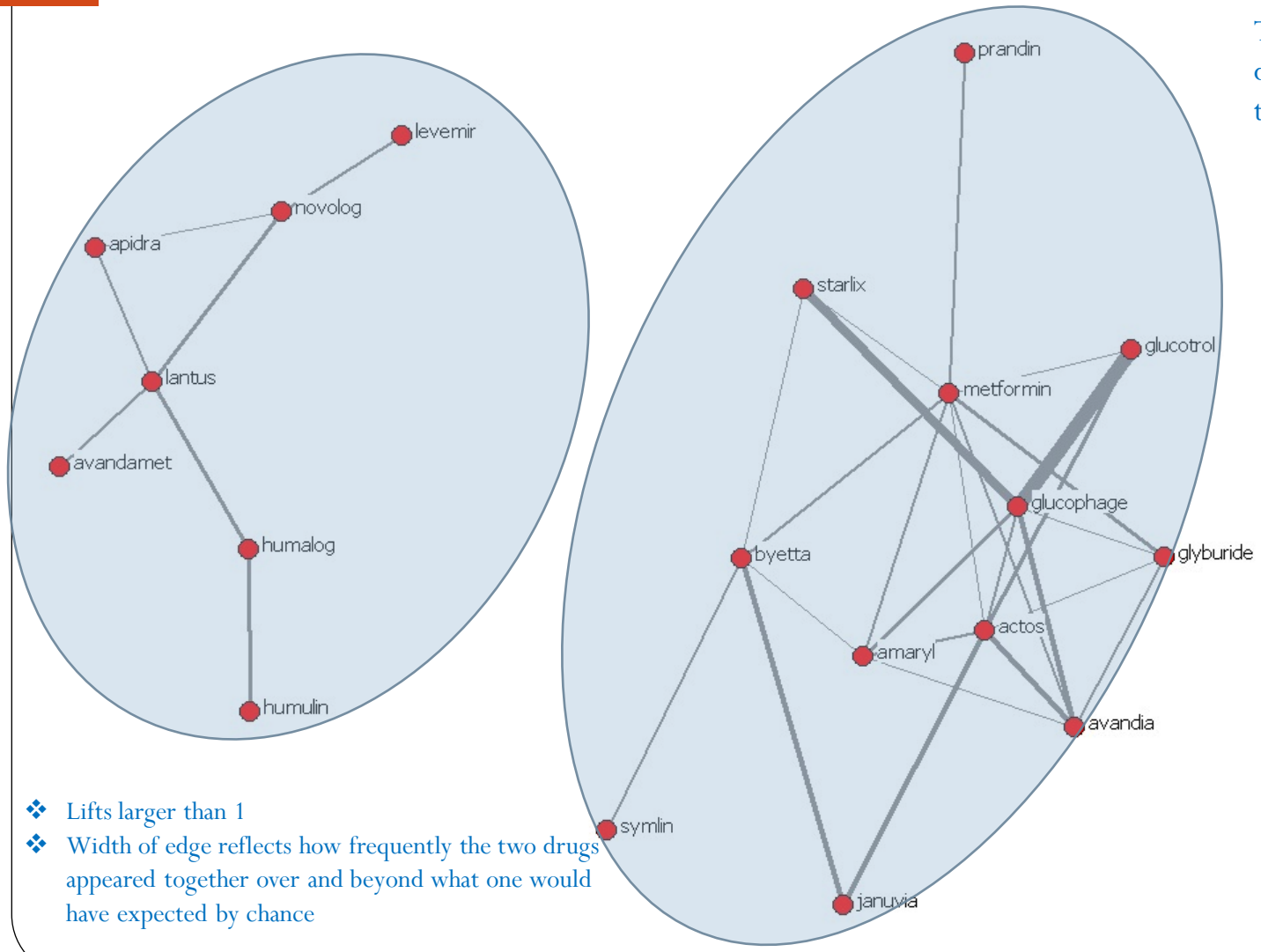
Byetta was mentioned frequently with:

- Glucotrol
- Januvia
- Amaryl
- Actos
- Avandia
- Prandin
- Symlin

- Lifts larger than 3
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Drug Usage Analysis

Drug Co-taking – Drugs mentioned as "Taken Together"



There are two main clusters of drugs that are mentioned as "taken together"

Byetta was mentioned as "taken together" with:

- Januvia
- Symlin
- Metformin
- Amaryl
- Starlix

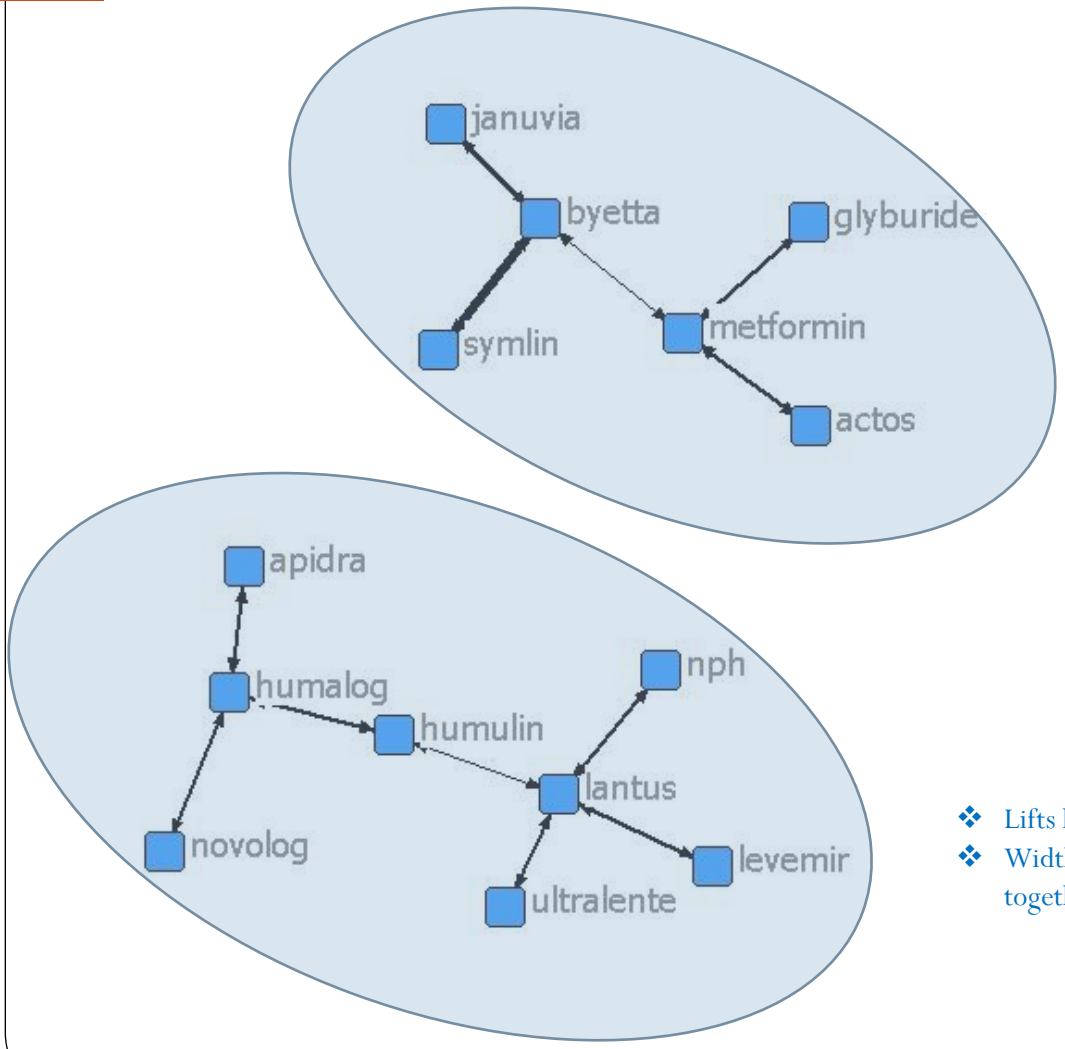
Pairs of drugs that are taken frequently together include:

- Glucotrol--Glucophage
- Glucophage--Stralix
- Byetta--Januvia
- Avandia--Actos
- Glucophage--Avandia

- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Drug Usage Analysis

Drug Switching – Drugs mentioned as "Switched" to and from



There are two main clusters of diabetes drugs within which consumers mentioned frequently that they "switched" from one drug to another

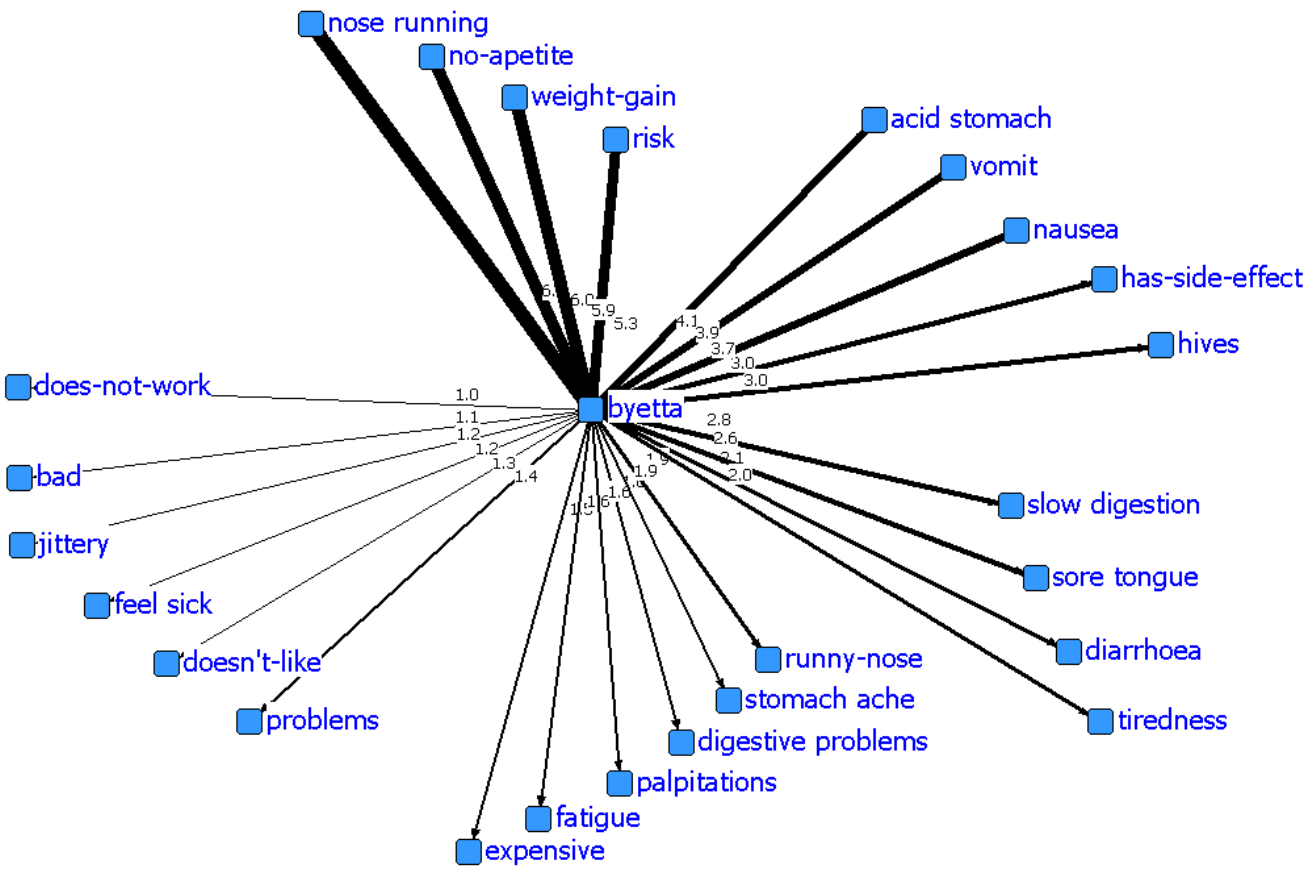
Byetta was mentioned as "switched" to and from:

- Symlin
- Januvia
- Metformin

- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Drug Terms Analysis

Byetta - Side Effects Analysis

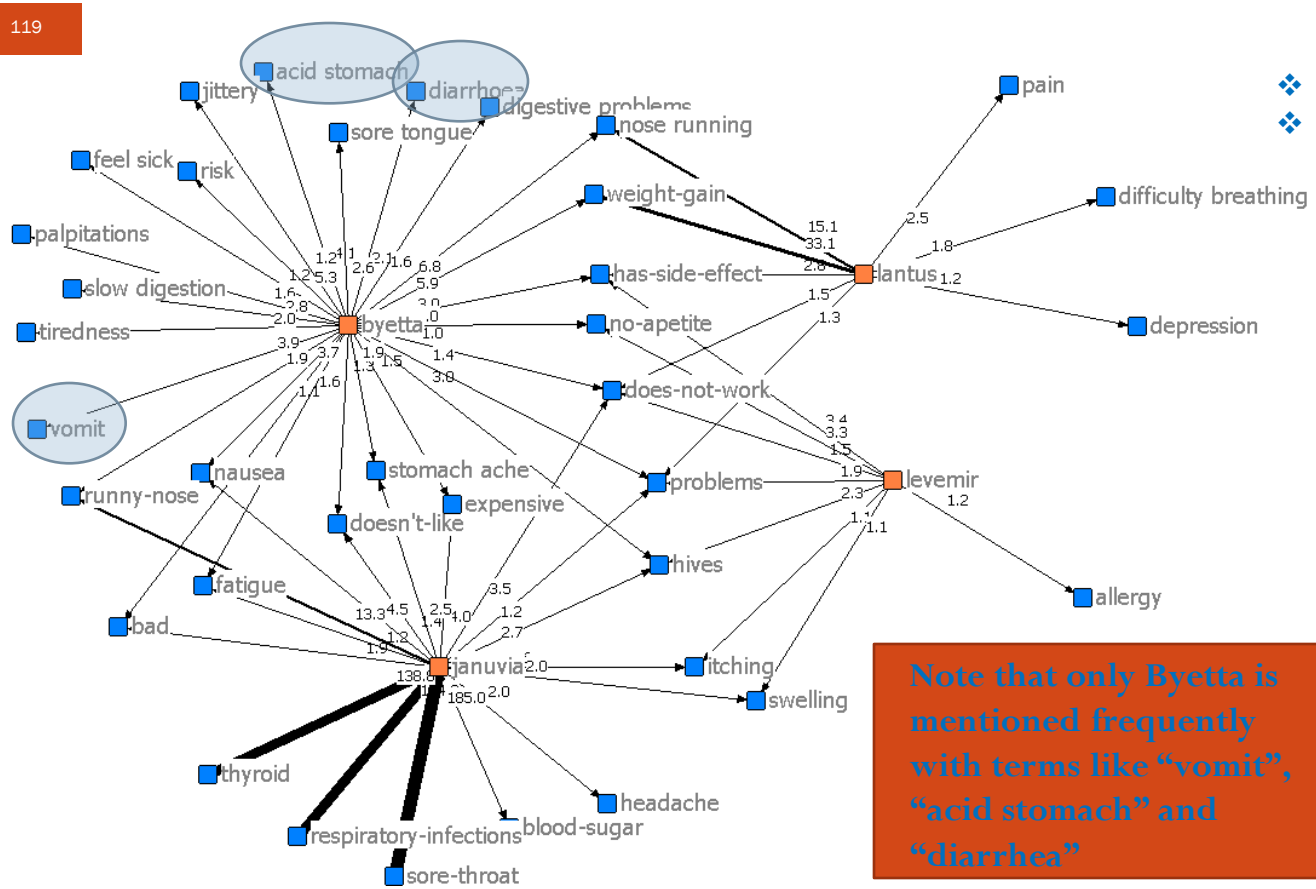


Byetta appeared much more than chance with the following side effects:

- "Nose running" or "runny nose"
- "No appetite"
- "Weight gain"
- "Acid stomach"
- "Vomit"
- "Nausea"
- "Hives"

Drug Terms Analysis

Drug Comparisons on Side Effects



- Lifts larger than 1
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

The main side effects discussed with Januvia:

- Thyroid
- Respiratory infections
- Sore throat

The main side effects discussed with Levemir :

- No appetite
- Hives

The main side effects discussed with Lantus:

- Weight gain
- Nose running
- Pain

Note that only Byetta is mentioned frequently with terms like "vomit", "acid stomach" and "diarrhea"

Byetta shares with Januvia the side effects:

- Runny nose
- Nausea
- Stomach ache
- Hives

Byetta shares with Levemir the side effects:

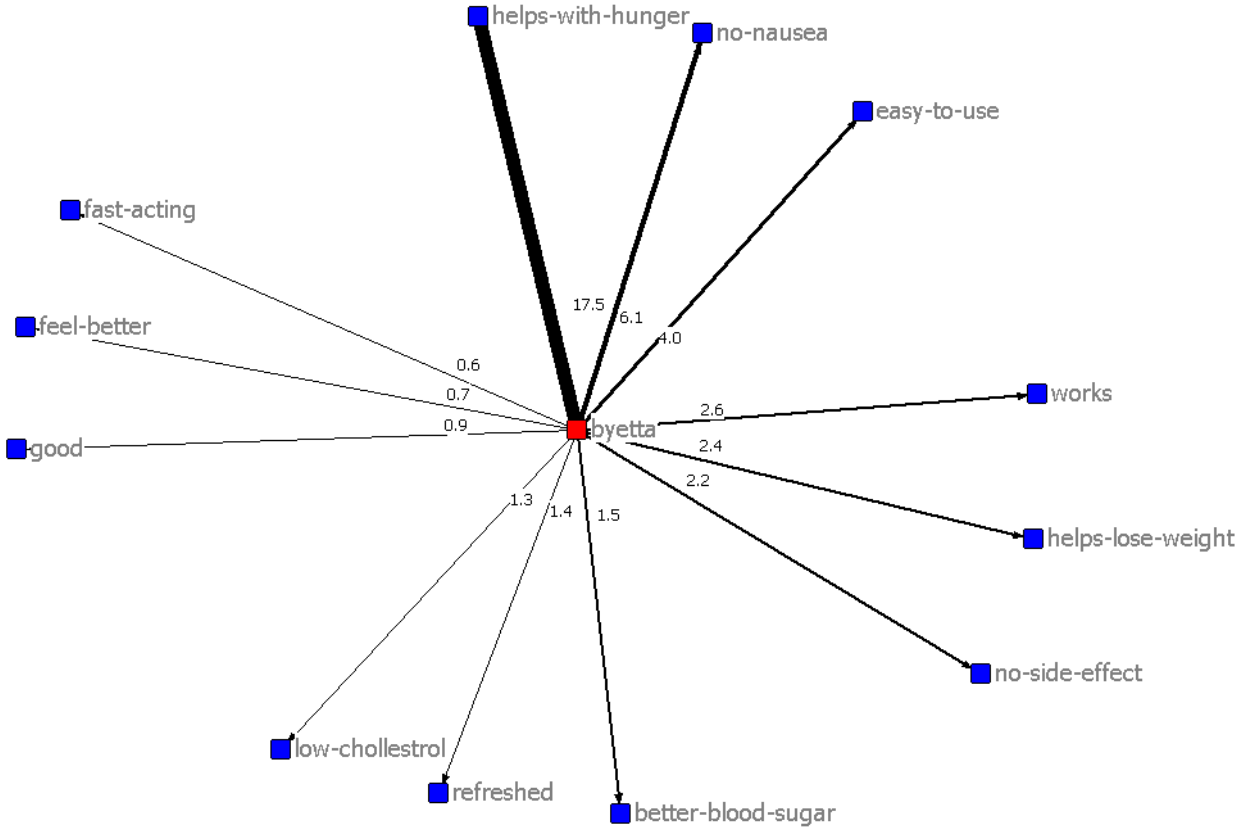
- No appetite
- Hives

Byetta shares with Lantus the side effects:

- Nose running
- Weight gain

Drug Terms Analysis

Byetta – Positive Sentiments

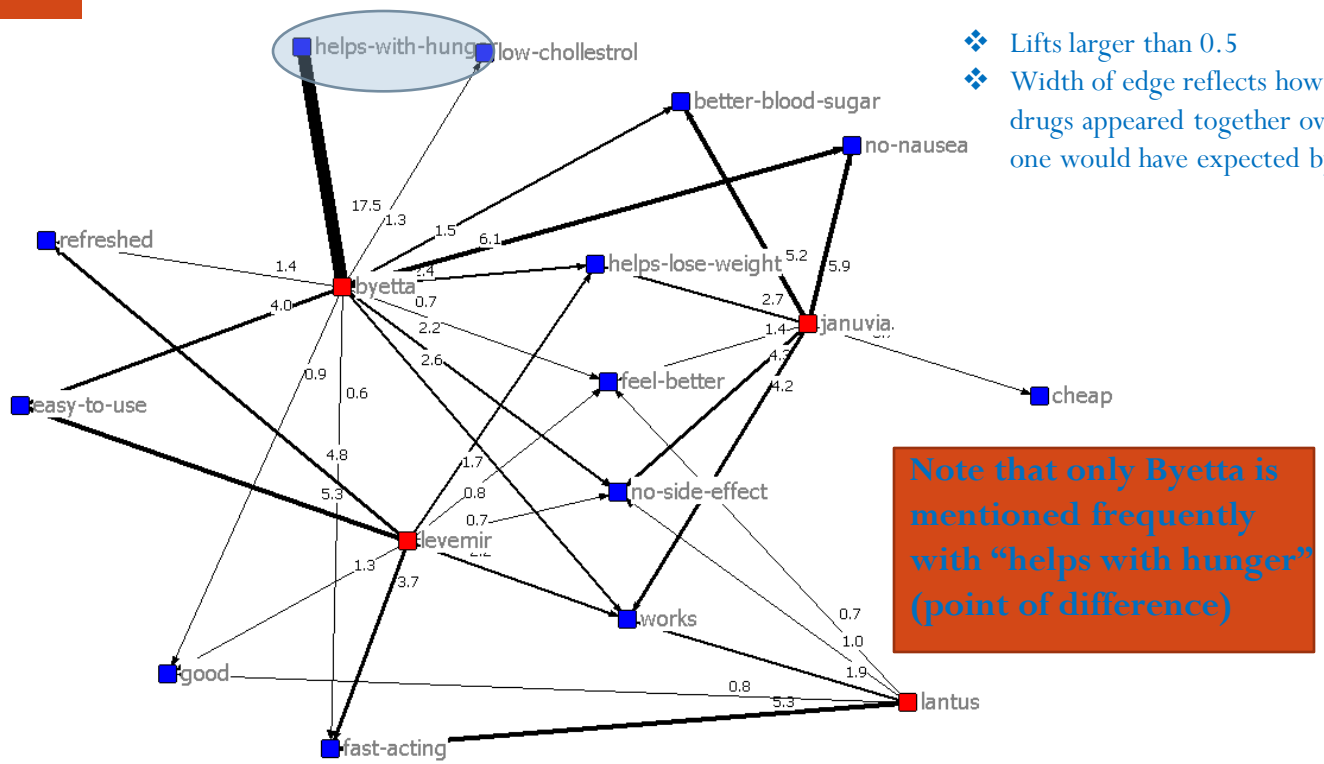


Byetta appeared much more than chance (lift>2) with the following positive sentiments:

- "Helps with hunger"
- "No nausea"
- "Easy to use"
- "Works"
- "Helps losing weight"
- "No side effects"

Drug Terms Analysis

Drug Comparisons on Positive Sentiments



- Lifts larger than 0.5
- Width of edge reflects how frequently the two drugs appeared together over and beyond what one would have expected by chance

Note that only Byetta is mentioned frequently with "helps with hunger" (point of difference)

The main positive sentiments discussed with Januvia:
- "No nausea"
- "Better blood sugar"
- "Works"
- "No side effects"

The main positive sentiments discussed with Levemir :
- "Easy to use"
- "Fast acting"

The main positive sentiments discussed with Lantus:
- "Fast acting"
- "Works"

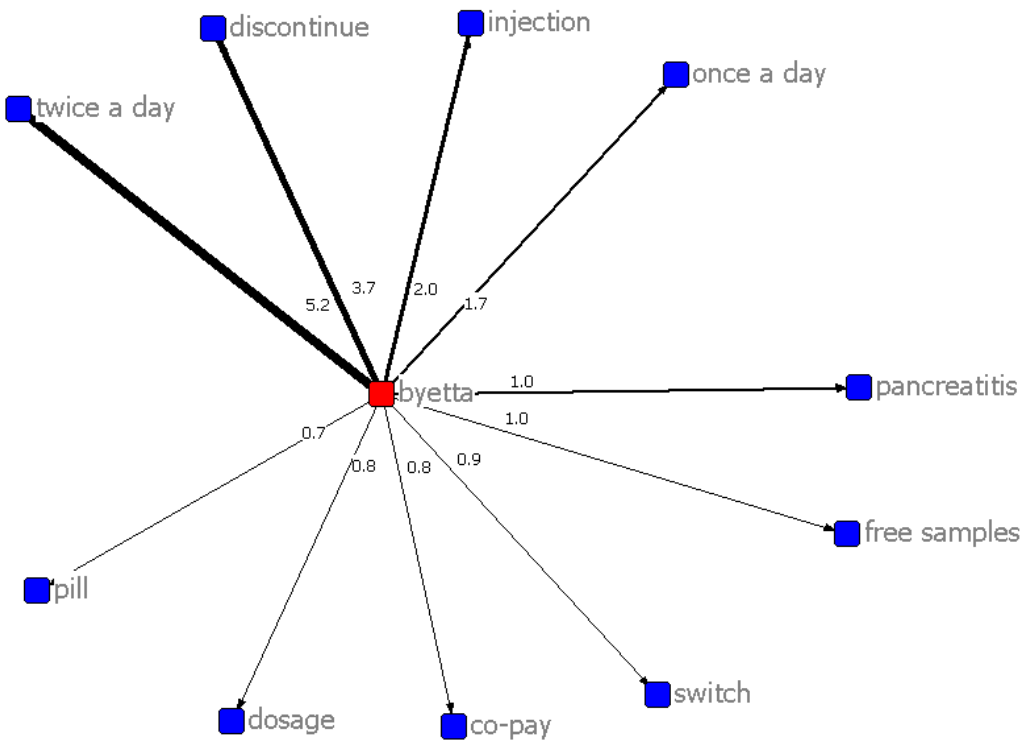
Byetta shares with Januvia:
- "Better blood sugar"
- "No nausea"
- "Helps lose weight"
- "No side effects"
- "Works"

Byetta shares with Levemir:
- "Easy to use"
- "Helps lose weight"
- "No side effects"
- "Works"

Byetta shares with Lantus:
- "Easy to use"
- "No side effects"
- "Works"

Drug Terms Analysis

Byetta – Other Sentiments



Byetta appeared much more than chance (lift>1.5) with the following sentiments:

- "Twice a day"
- "Discontinue"
- "Injection"
- "Once a day"

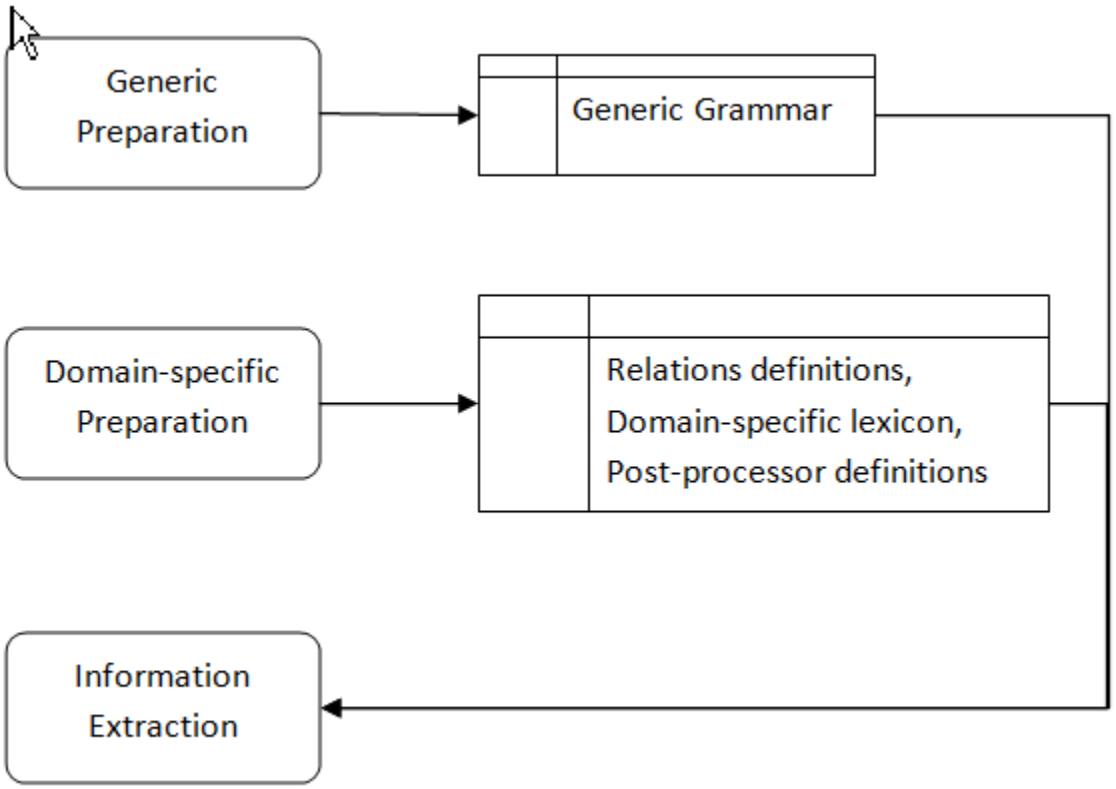
Byetta was mentioned moderately with the sentiments:

- "Pancreatitis"
- "Free sample"

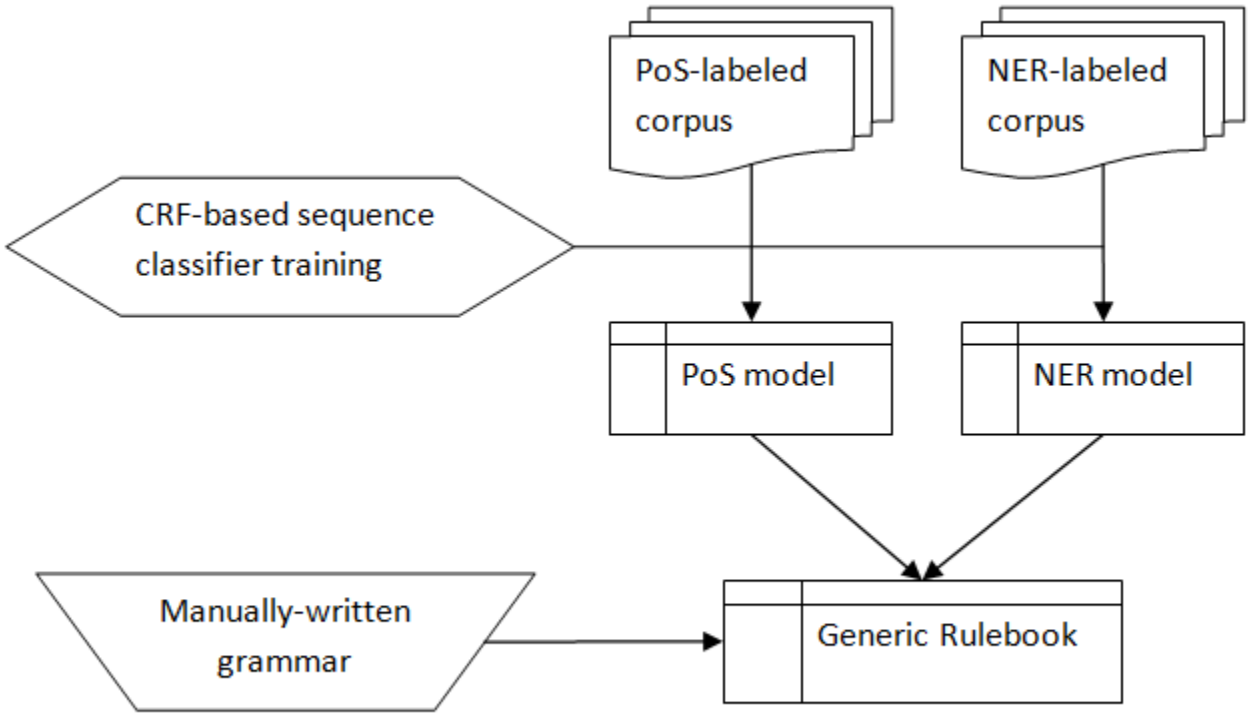
Visual care

IE authoring environment

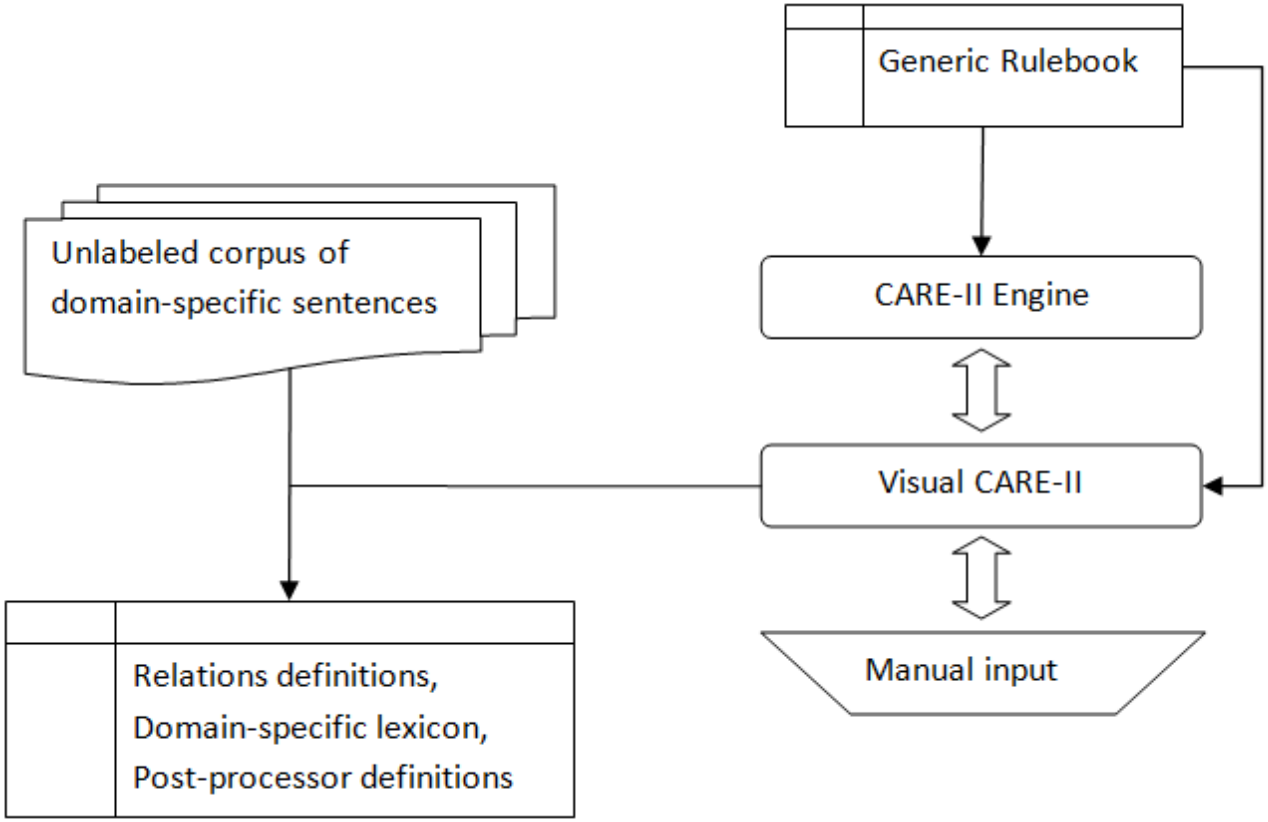
Overall architecture



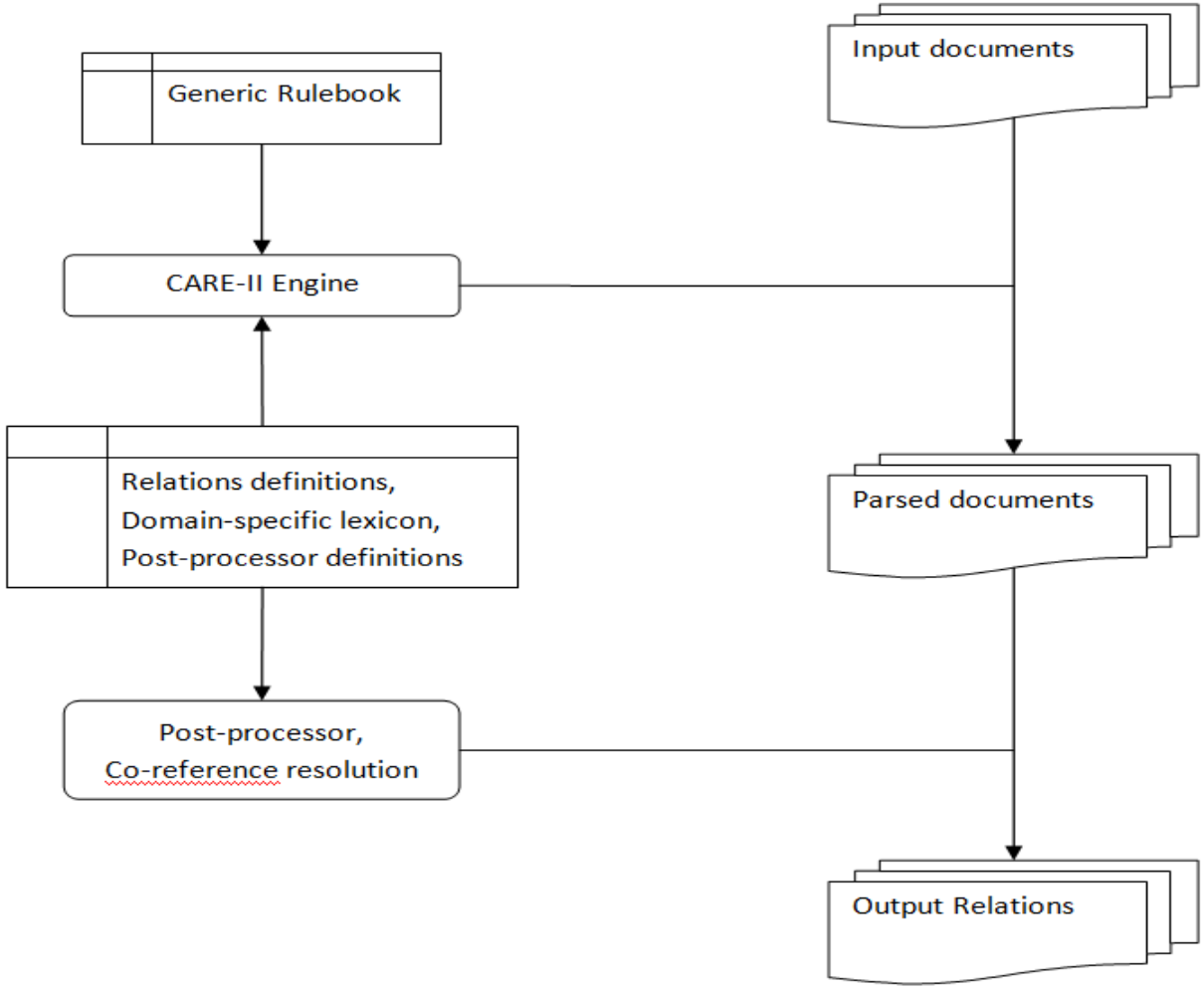
Generic preparation stage



Domain specific preparation stage



Information extraction stage



Wrap-up

What did we cover?

- Intro to text mining
  - Information *Retrieval* (IR) vs. Information *Extraction* (IE)
- Information extraction (IE)
  - IE Components
- Open IE/Relation Extraction
  - Basic Open IE: Text Runner
  - Advanced Open IE: KnowItAll and SRES
- Sentiment mining
  - its relation to IE
- Visualization of text mining results
  - What is compared to what, and how
- Wrap-up

Text Mining is Big Business

- **Part of most big data mining systems**
  - SAS, Oracle, SPSS, SAP Fair Isaac, …
- ***Many* sentiment analysis companies**
  - the "big boys," Nielsen Buzzmetrics and dozens of others.
- **Sometimes tied to special applications**
  - **Autonomy** - suite of text mining, clustering and categorization solutions for knowledge management
  - **Thomson Data Analyzer** - analysis of patent information, scientific publications and news
- **Open source has more fragmented tools**
  - **NLTK, Stanford NLP tools, GATE, Lucene, MinorThird**
  - **RapidMiner/YALE** - open-source data and text mining
- **Lots more:**
  - **AeroText** - Information extraction in multiple languages
  - **LanguageWare** - the IBM Tools for Text Mining.
  - Attensity, Endeca Technologies, Expert System S.p.A., Nstein Technologies

Summary

- Information Extraction
  - Not just information retrieval
  - Find named entities, relations, events
  - Hand-built vs. trained models
    - CRFs widely used
- Open Information Extraction
  - Unsupervised relation extraction
    - Bootstrap pattern learning
- Sentiment analysis
- Visualize results
  - Link analysis, MDS, …
- Text mining is easy and hard

To learn more

- See www.cis.upenn.edu/~ungar/KDD/text-mining.html